

# NGS data analysis

Tomáš Hron  
thron@hpst.cz



Autorizovaný distributor Agilent Technologies



Produkty

Služby

E-shop pro molekulární biologii

Kontakty

Servisní požadavek

O nás

Kariéra

Knihovna Agilent

Odkazy

Novinky

Semináře

Novinky a trendy Agilent  
Technologies 15. 10. 2020

Přehled HPST webinářů  
(aktuální i ze záznamu)

Setkání uživatelů platformy  
VarSome Clinical

Rozrostli jsme se o  
**spotřební materiál  
a laboratorní vybavení**  
[www.labicom.cz](http://www.labicom.cz) | [eshop.labicom.cz](http://eshop.labicom.cz)

Labicom + hpst → HPST  
**Fúze společností HPST, s.r.o.  
a Labicom s.r.o.**



Agilent GC, GC-MS and GC-MS/MS  
Recorded Webinars

GC & GC/MS Webinars



NOVINKA v naší nabídce - protilátky a reagencie DAKO pro  
průtokovou cytometrii

Agilent  
**Dako**

Protilátky a reagencie  
DAKO pro průtokovou  
cytometrii

# Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis

# Program přednášky

- **Bioinformatická analýza**
- Galaxy
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis

# Ten simple rules for providing effective bioinformatics research support

Judit Kumuthini , Michael Chimenti, Sven Nahnsen, Alexander Peltzer, Rebone Meraba, Ross McFadyen, Gordon Wells, Deanne Taylor, Mark Maienschein-Cline, Jian-Liang Li, Jyothi Thimmapuram, Radha Murthy-Karuturi, Lyndon Zass

Published: March 26, 2020 • <https://doi.org/10.1371/journal.pcbi.1007531>

## Project development

**Rule 1: Collaboratively design experiment**

**Rule 2: Manage scope and expectations**

**Rule 3: Define and ensure data management**

## Data collection and generation

**Rule 4: Manage the traceability of data**

**Rule 5: Determine how and what metadata are reported**

**Rule 6: Coordinate data and internet security**

## Data analysis

**Rule 7: Control data quality throughout the project lifecycle**

**Rule 8: Identify suitable computational tools for data analysis**

**Rule 9: Track, record, and confirm workflow changes**

**Rule 10: Repurpose the data**



"Data science done well looks easy – and that's a big problem for data scientists."

Jeff Leek

Associate Professor at Johns Hopkins Bloomberg  
School of Public Health.



HPST, s.r.o.

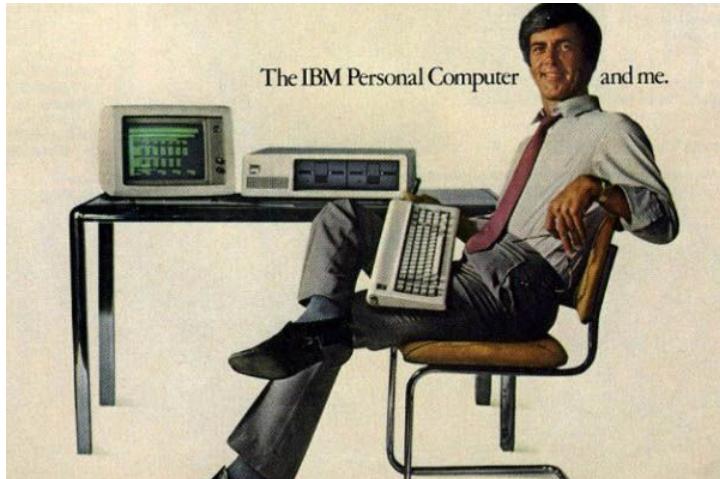
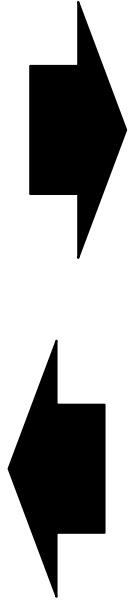


# Remote / cloud computing

“There is no reason anyone would want a computer in their home.”

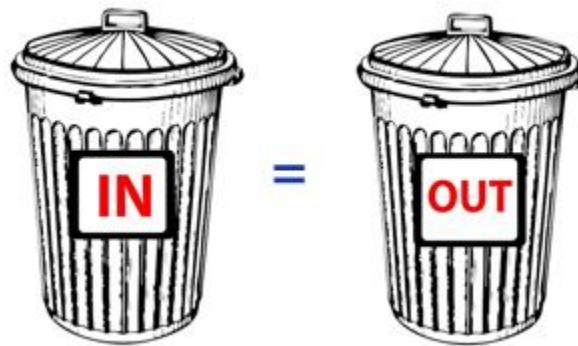
- Ken Olsen, 1977

Co-founder of Digital Equipment Corporation (DEC)



# Dobra data jsou pedpokladem hladke analyzy a dobreho vsledku

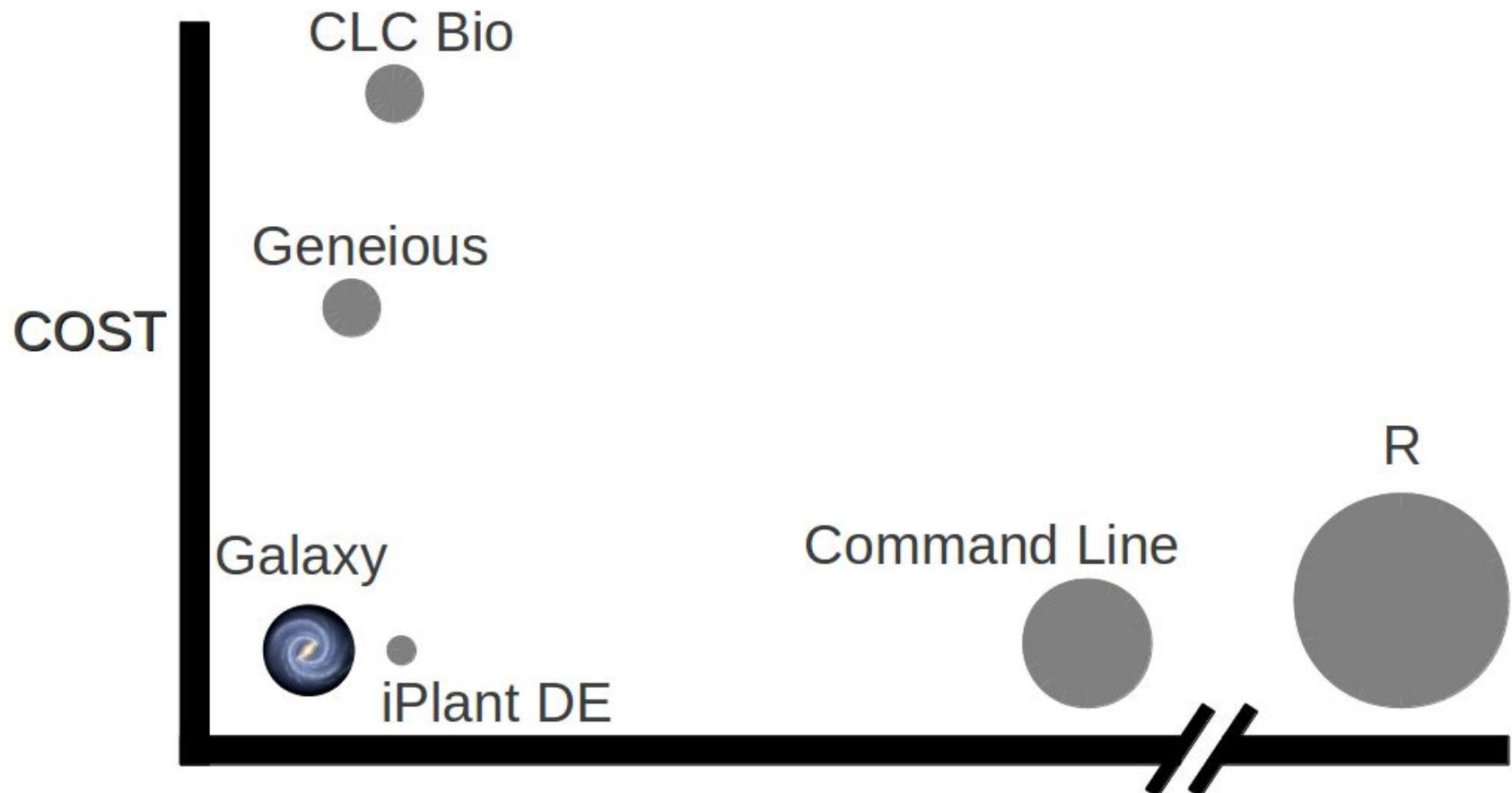
**Garbage in → Garbage out**



# Program přednášky

- Bioinformatická analýza
- **Galaxy**
- NGS
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis





Size of dot indicates flexibility/power

# Co je Galaxy?



“Galaxy je otevřená webová platforma pro analýzu biomedicínských dat.”

- Původně byla vytvořena pro genomiku, ale v současnosti obsahuje nástroje pro bioinformatickou analýzu obecně

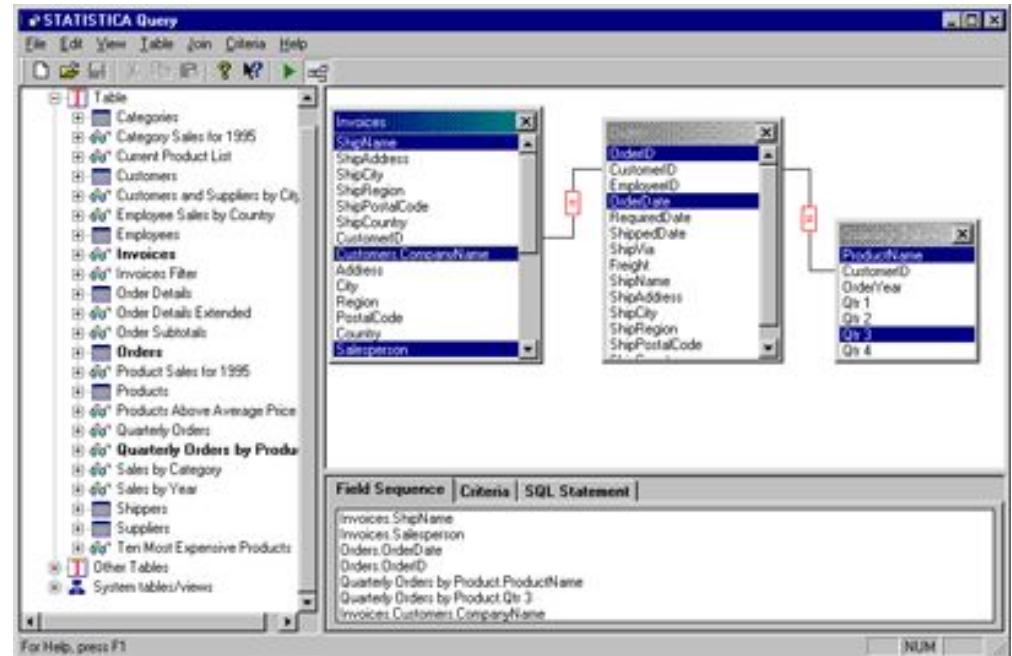
# Command Line Interface

## vs

# Graphical User Interface

“Graphical user interfaces make easy tasks easy, while command line interfaces make difficult tasks possible”

```
ja@muj-VirtualBox:~$ ls -l
total 44
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Desktop
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Documents
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Downloads
-rw-r--r-- 1 ja ja 8980 kvě 26 21:52 examples.desktop
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Music
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Pictures
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Public
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Templates
drwxr-xr-x 2 ja ja 4096 kvě 26 21:57 Videos
ja@muj-VirtualBox:~$ history
 1 cat /proc/cpuinfo
 2 python
 3 sudo apt-get install python3
 4 python3
 5 sudo apt-get install samtools
 6 sudo apt-get install vcftools
 7 vcftools
 8 samtools
 9 ls -la
10 clear
11 ls -l
12 history
ja@muj-VirtualBox:~$
```



# odstranění nekvalitních čtení v Linuxu

```
thron@ntb-thron:~$  
thron@ntb-thron:~$ java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 6 -phred33 input_R1.fa  
stq input_R2.fastq output_R1.fq output_U1.fq output_R2.fq output_U2.fq AVGQUAL:20
```



# odstranění nekvalitních čtení v galaxy

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.3)

**Single-end or paired-end reads?**

Single-end

**Input FASTQ file**

No fastqsanger or fastqsanger.gz dataset available.

**Perform initial ILLUMINACLIP step?**

Yes No

Cut adapter and other illumina-specific sequences from the read

**Trimomatic Operation**

**1: Trimmomatic Operation**

**Select Trimmomatic operation to perform**

Drop reads with average quality lower than a specified level (AVGQUAL)

**Minimum average quality required to keep a read**

20

**Insert Trimmomatic Operation**

**Execute**



HPST, s.r.o.



# Jak dlohuje existuje Galaxy?

Od roku 2005

## Galaxy: A platform for interactive large-scale genome analysis

Belinda Giardine,<sup>1</sup> Cathy Riemer,<sup>1</sup> Ross C. Hardison,<sup>1</sup> Richard Burhans,<sup>1</sup> Laura Elnitski,<sup>2</sup> Prachi Shah,<sup>1,2</sup> Yi Zhang,<sup>1</sup> Daniel Blankenberg,<sup>1</sup> Istvan Albert,<sup>1</sup> James Taylor,<sup>1</sup> Webb Miller,<sup>1</sup> W. James Kent,<sup>3</sup> and Anton Nekrutenko<sup>1,4</sup>

<sup>1</sup>Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA; <sup>2</sup>National Human Genome Research Institute, Bethesda, Maryland 20892, USA; <sup>3</sup>Department of Computer Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA

Accessing and analyzing the exponentially expanding genomic sequence and functional data pose a challenge for biomedical researchers. Here we describe an interactive system, Galaxy, that combines the power of existing genome annotation databases with a simple Web portal to enable users to search remote resources, combine data from independent queries, and visualize the results. The heart of Galaxy is a flexible history system that stores the queries from each user; performs operations such as intersections, unions, and subtractions; and links to other computational tools. Galaxy can be accessed at <http://g2.bx.psu.edu>.

[Supplemental material is available online at [www.genome.org](http://www.genome.org).]

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240089/pdf/00151451.pdf>

# Kdo vytváří Galaxy?



PennState

Center for Comparative Genomics and Bioinformatics



JOHNS HOPKINS  
UNIVERSITY

Department of Biology

A velké množství přispěvatelů z komunity Galaxy...



HPST, s.r.o.

Agilent Technologies  
Autorizovaný distributor

# Jak začít používat Galaxy?

- Na veřejném serveru **usegalaxy.org**
- Pro akademiky a studenty na serveru Metacentra
- Instalace na serveru vaší laboratoře

# Cesnet a MetaCentum

bezplatné využití výpočetní a úložné kapacity a řady  
aplikačních programů pro akademiky



HPST, s.r.o.



# Důležité odkazy

## Cesnet a MetaCentum

<https://metavo.metacentrum.cz/> - MetaCentrum VO

[https://wiki.metacentrum.cz/wiki/Main\\_Page](https://wiki.metacentrum.cz/wiki/Main_Page) - wiki stránka MetaCentra

[https://wiki.metacentrum.cz/wiki/Galaxy\\_application](https://wiki.metacentrum.cz/wiki/Galaxy_application) - wiki Galaxy v MetaCentru

<https://galaxy.metacentrum.cz/> - přístup na MetaCentrum Galaxy

## Galaxy

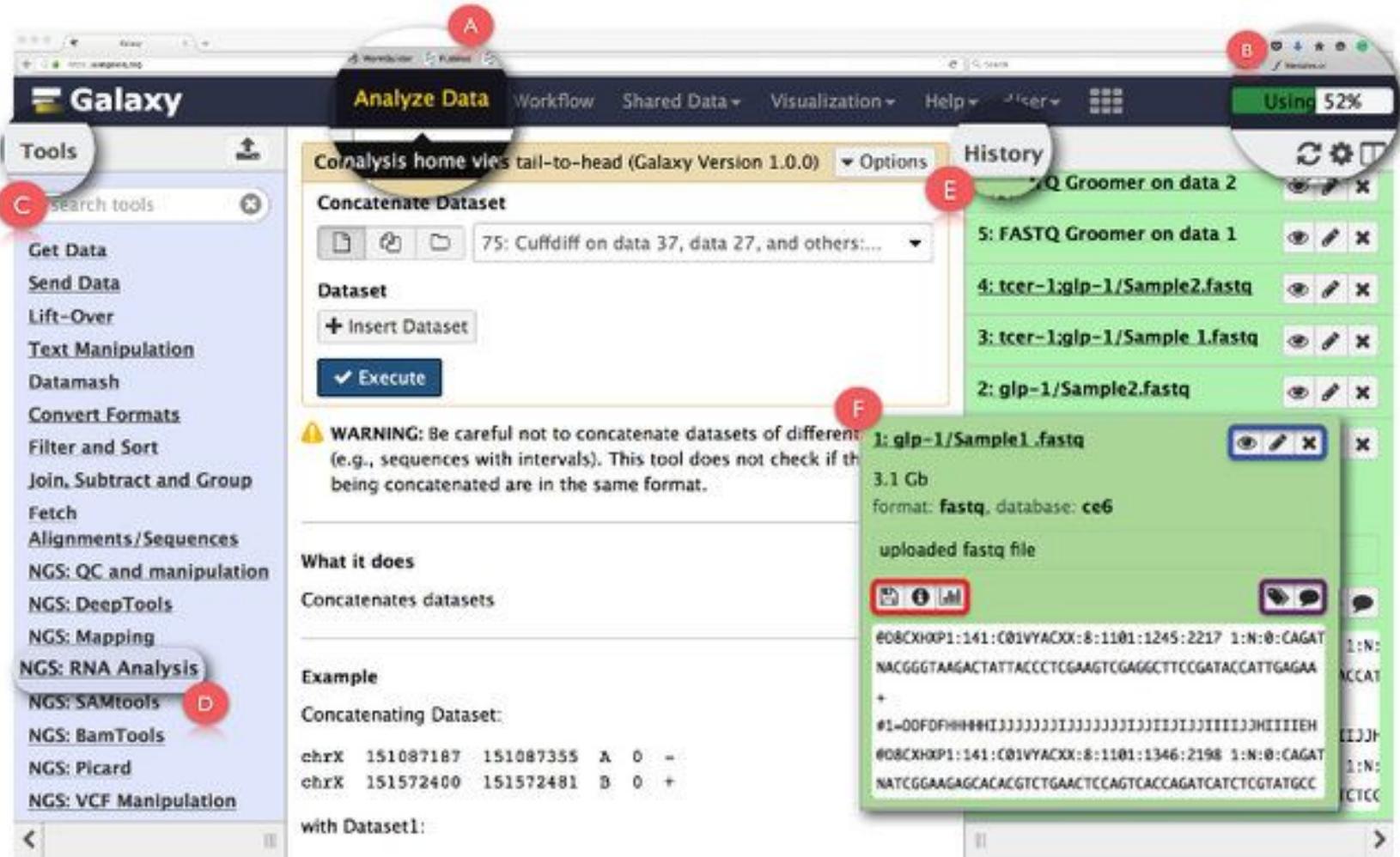
<https://wiki.galaxyproject.org/Learn> - Galaxy Wiki

<https://vimeo.com/galaxyproject> - Video tutoriály

[http://wiki.bits.vib.be/index.php/Galaxy\\_beginner's\\_tutorial](http://wiki.bits.vib.be/index.php/Galaxy_beginner's_tutorial) - Galaxy tutorial

<https://www.coursera.org/learn/galaxy-project> - Galaxy Coursera

<https://biostar.usegalaxy.org/> - Galaxy Biostars

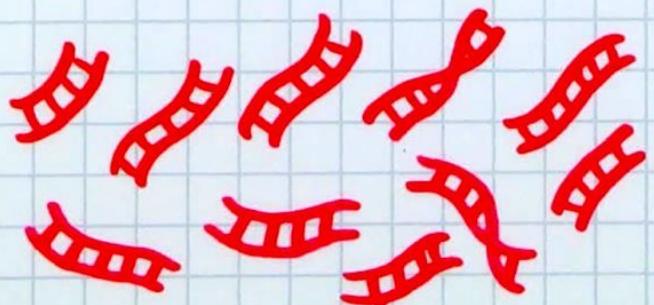
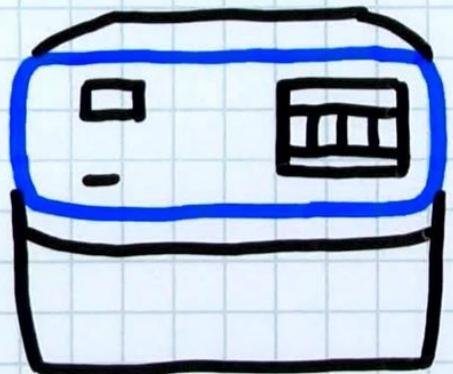


(A) highlights the 'Analyze data' function in the webpage header used to access Analysis Home View. (B) is the 'Progress bar' that indicates the space on the Galaxy server utilized by the operation. (C) is the 'Tools Section' that lists all the tools that can be run on the Galaxy interface. (D) shows the 'NGS: RNA Analysis' tool section used for RNA-Seq analysis. (E) depicts the 'History' panel that lists all the files generated using Galaxy. (F) shows an example of the dialogue box that opens up when clicking on any file in the History section. Within (F), the blue box highlights icons that can be used to view, edit the attributes or delete the dataset, the purple box highlights icons that can be used to 'edit' the dataset tags or annotation, and, the red box indicates icons to download the data, view details of the task performed or rerun the operation. Please click [here](#) to view a larger version of this figure.

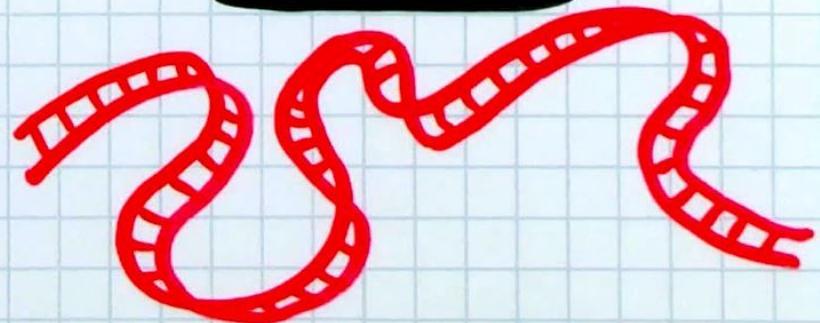
# Program přednášky

- Bioinformatická analýza
- Galaxy
- **NGS**
- Analysis Workflows
- Praktická část
  - KIR genotyping
  - RNAseq analysis

**NGS**  
MASSIVELY  
PARALLEL



**SANGER**



# NGS (Next Generation Sequencing)

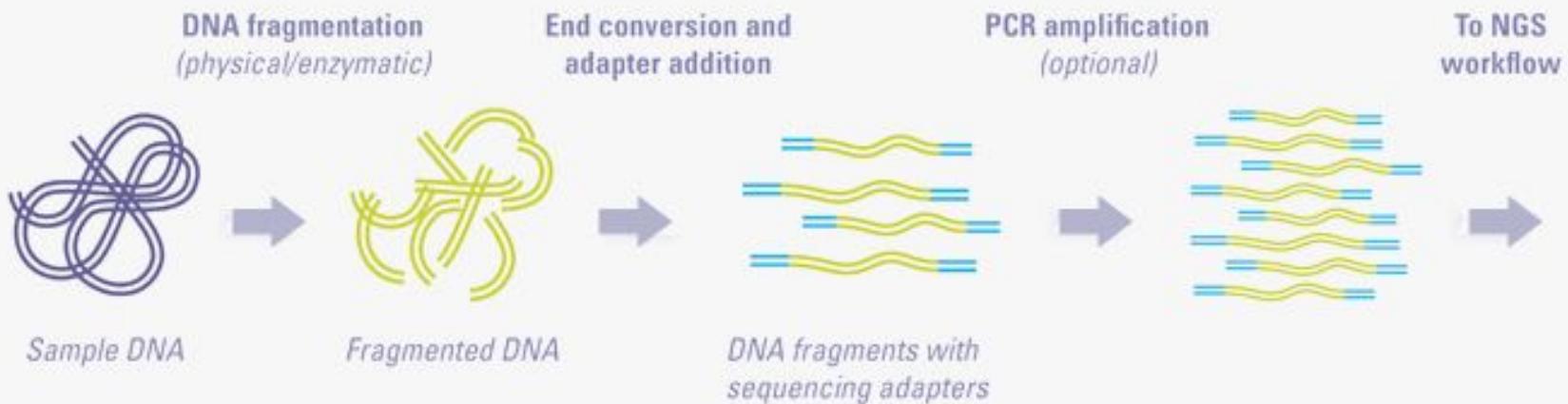
- Velké množství dat: od několika Gb po Tb
  - Vyžaduje automatizované zpracování dat
  - Vyžaduje infrastrukturu pro uskladnění dat
- Pokrytí: od několika genů po celý genom/transkriptom
- Vysoká citlivost ale náročnější interpretace pro klinické použití
- Volba velkého množství nástrojů pro zpracování dat

# Illumina

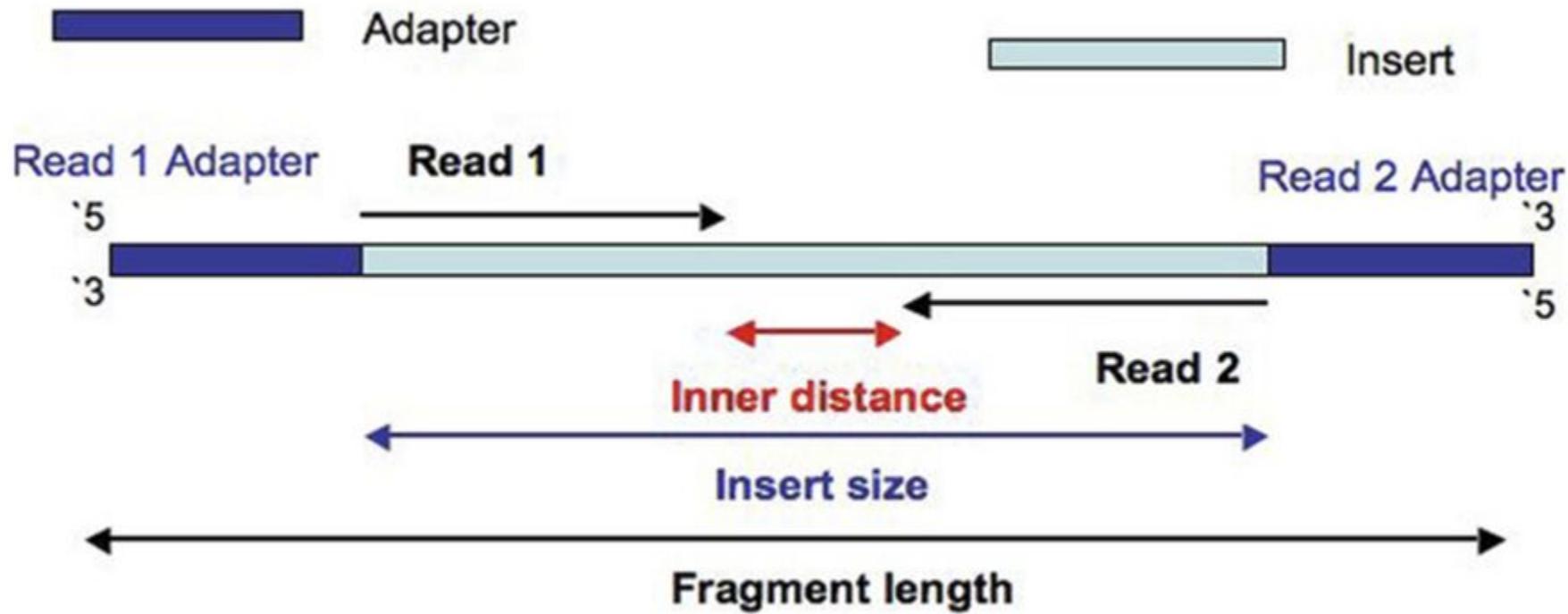
- 4M - 400M reads per run
- Small read size (50 - 600bp)
- Accuracy 0.1 - 1%
- Whole genome sequencing or target enrichment



# Sequencing libraries

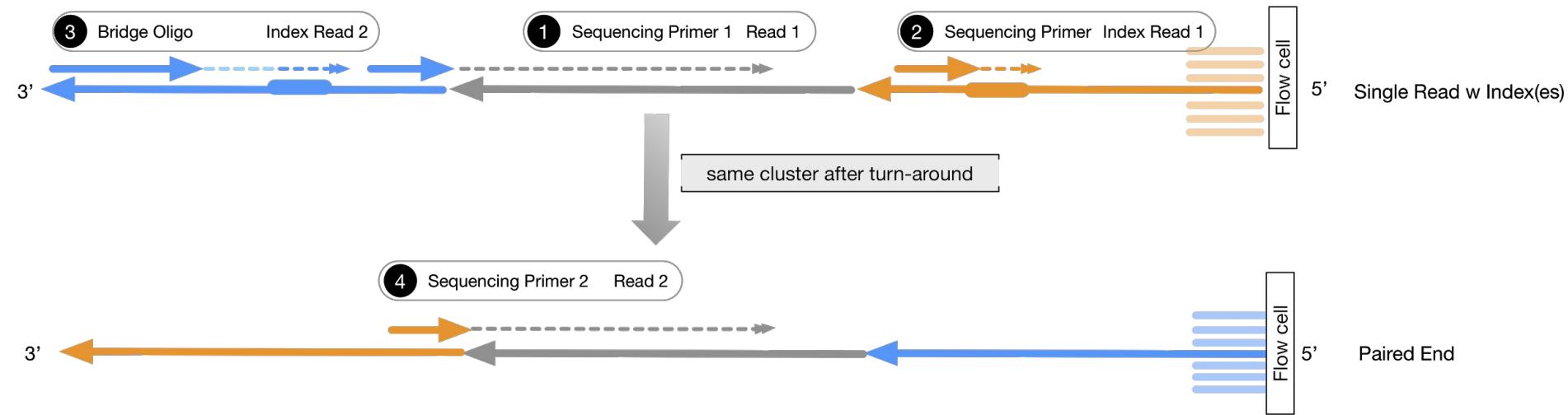


# Sequencing libraries



# Sequencing libraries

## Indexes - scheme of sequencing procedure



# DNA sequencing libraries

**Whole-genome lib.**

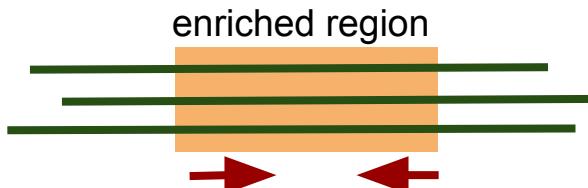


fragmentation  
of DNA



Sequencing

**Amplicon lib.**

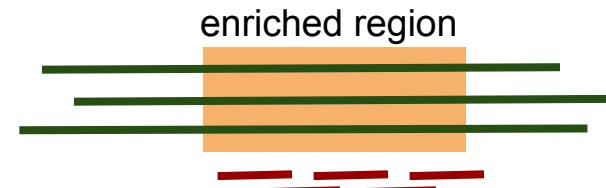


enrichment by  
specific primers

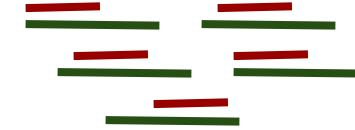


Sequencing

**Sequence capture lib.**

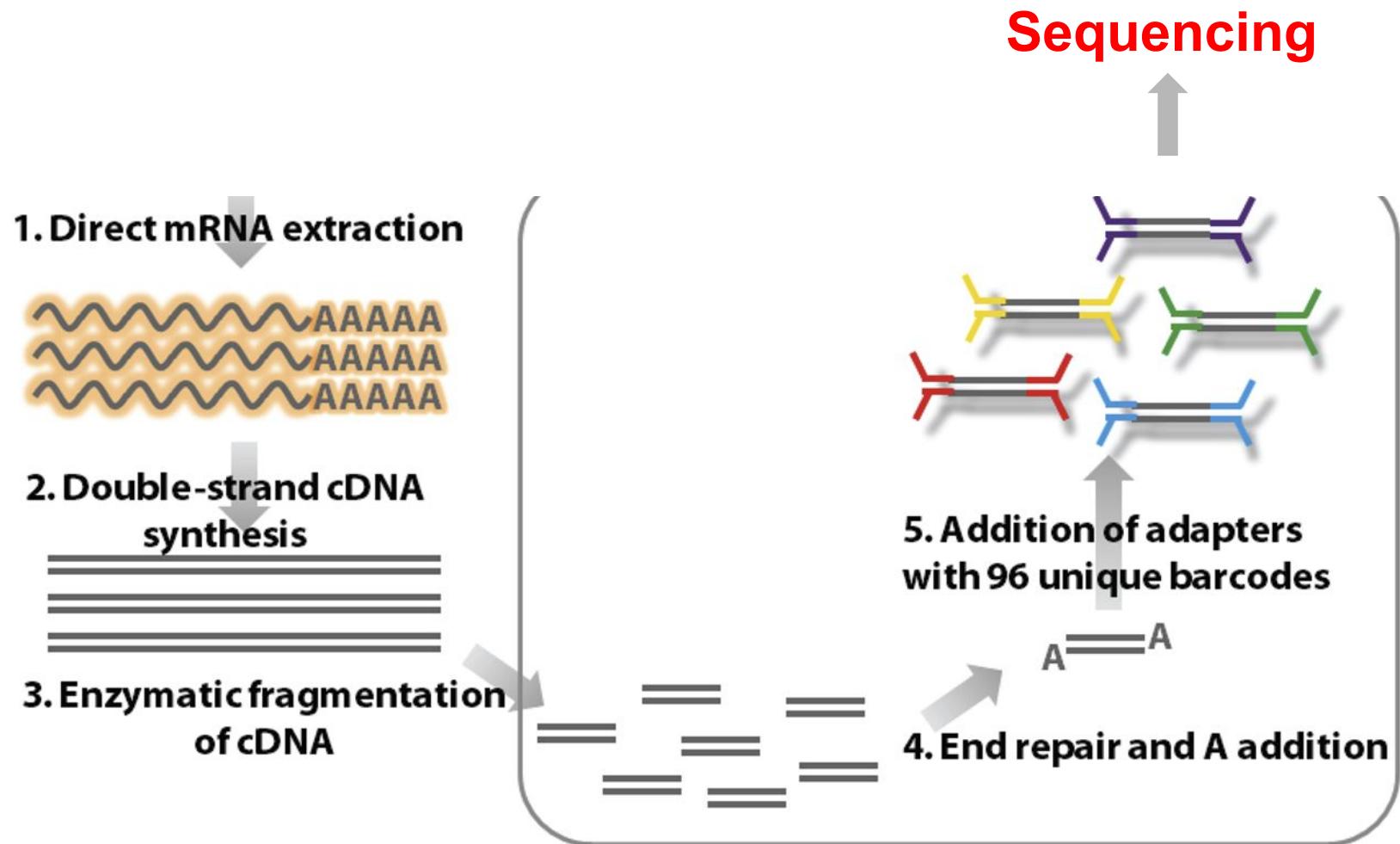


enrichment by  
specific probes



Sequencing

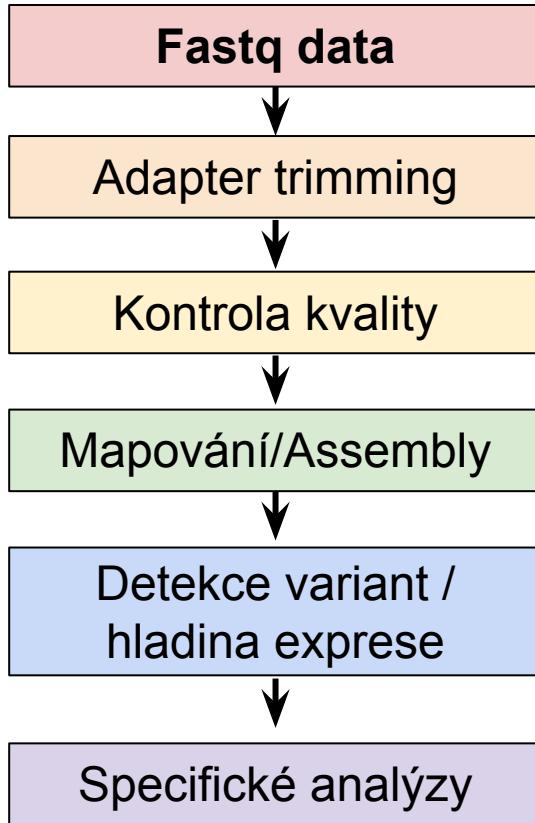
# RNA sequencing libraries



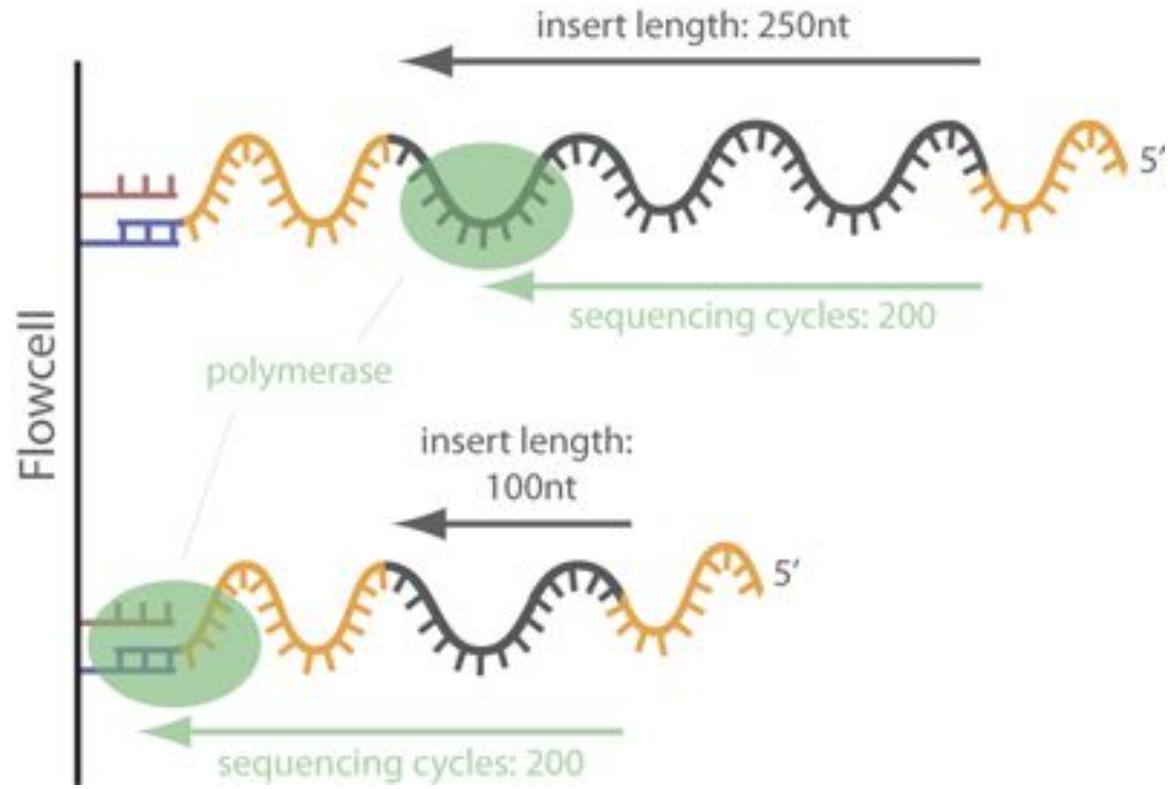
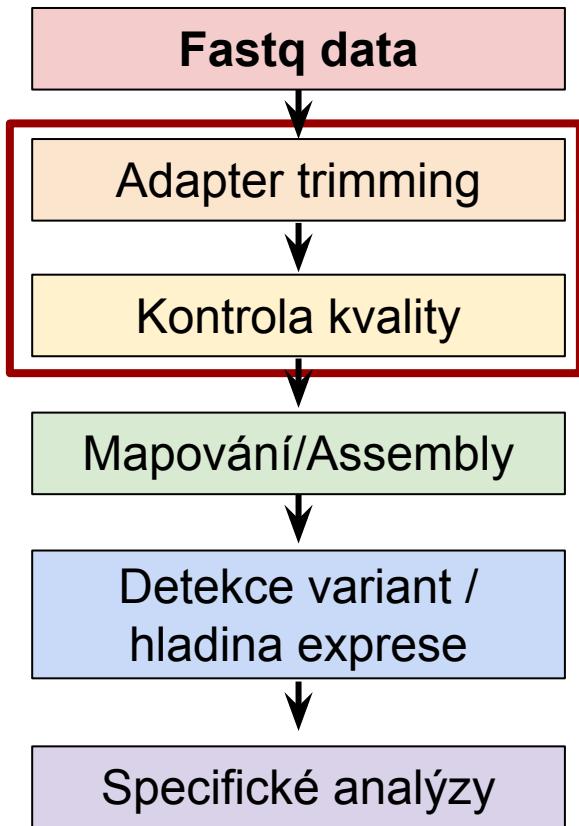
# Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- **Analysis Workflows**
- Praktická část
  - KIR genotyping
  - RNAseq analysis

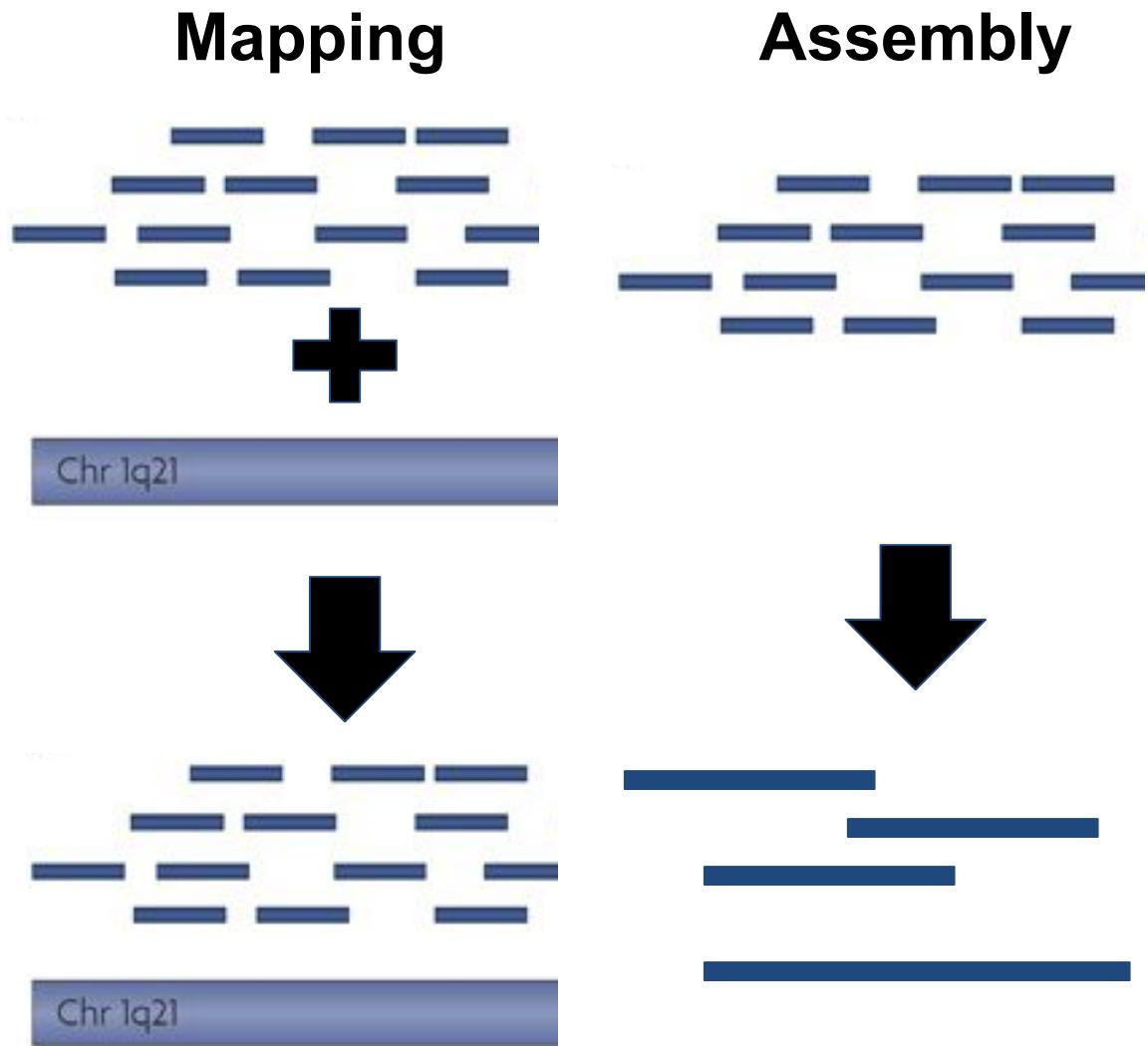
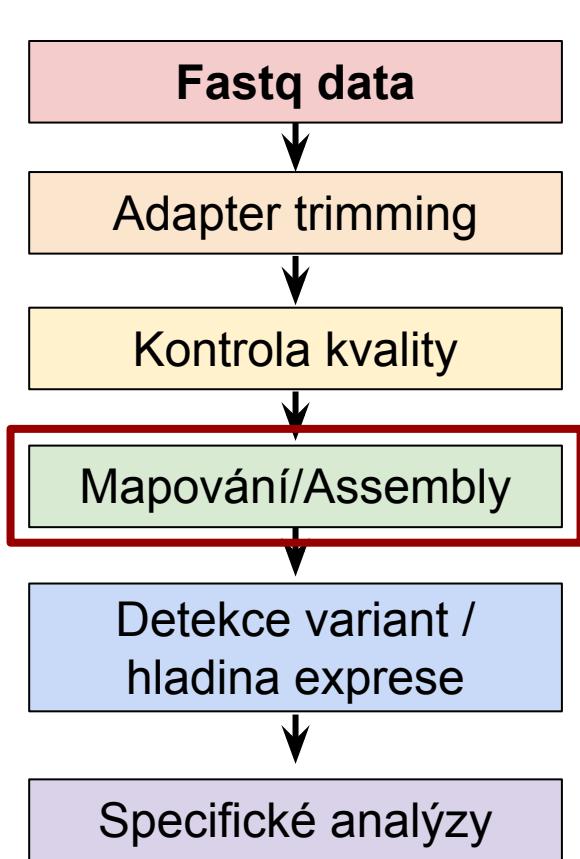
# Analysis workflows



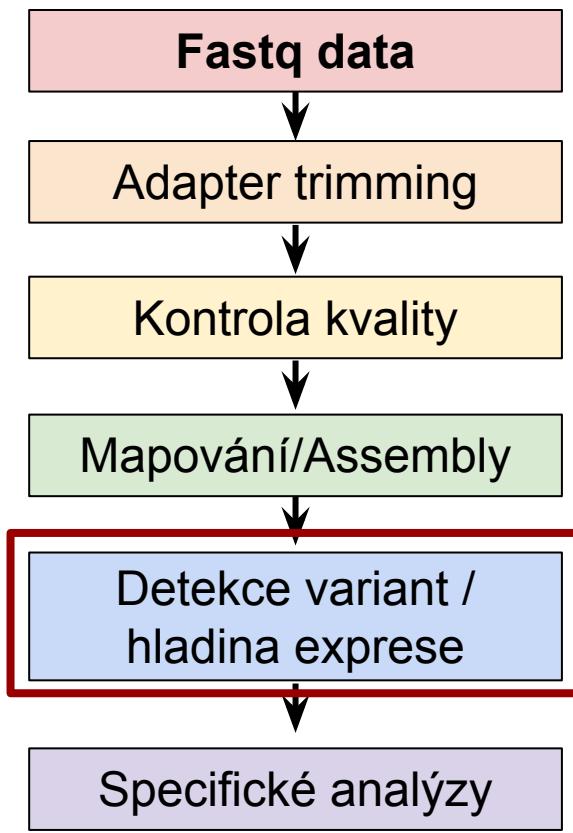
# Analysis workflows



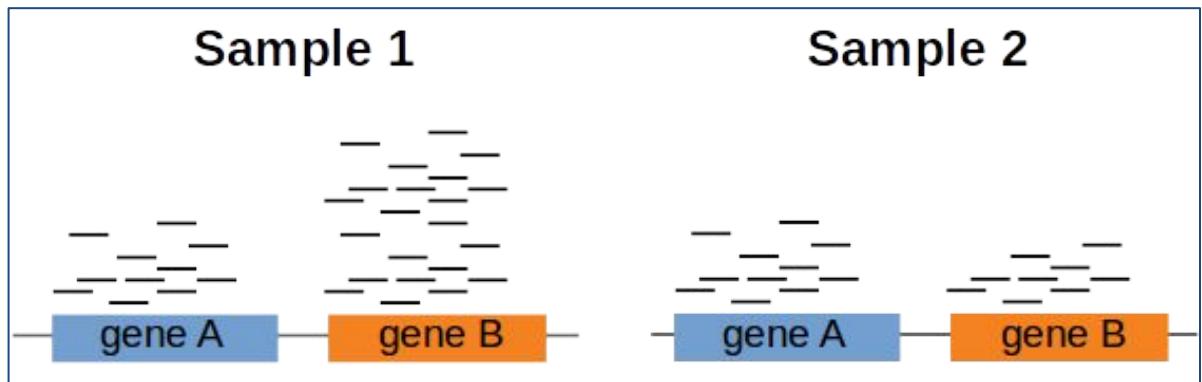
# Analysis workflows



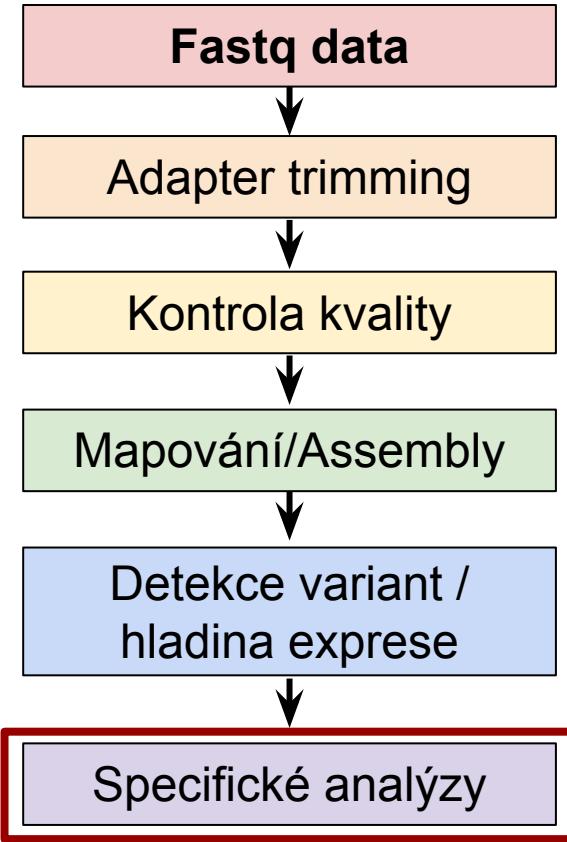
# Analysis workflows



Reference	Aligned Reads
	ACGATATTACACGTACACTCAAGTCGTTGGAACCT ACGATATTACACGTACA <b>TTC</b> AA <b>A</b> TCGT <b>ACG</b> ATATTACACGTACA <b>TT</b> CAA <b>C</b> TCGT ACGATATTACACGCACA <b>TT</b> CAAGTCGT CGAT <b>A</b> TTACACGTACA <b>TT</b> CAAGTCGTT ATATT <b>T</b> ACGTACA <b>TT</b> CAAGTCGTT <b>CG</b> ATATTAAAC <b>G</b> TACA <b>TT</b> CAAGTCGTT <b>CG</b> ATTACACGTACA <b>TT</b> CAAGTCGT <b>TC</b> GA ATTACACGTACA <b>TT</b> CACGT <b>CG</b> TT <b>CG</b> GA

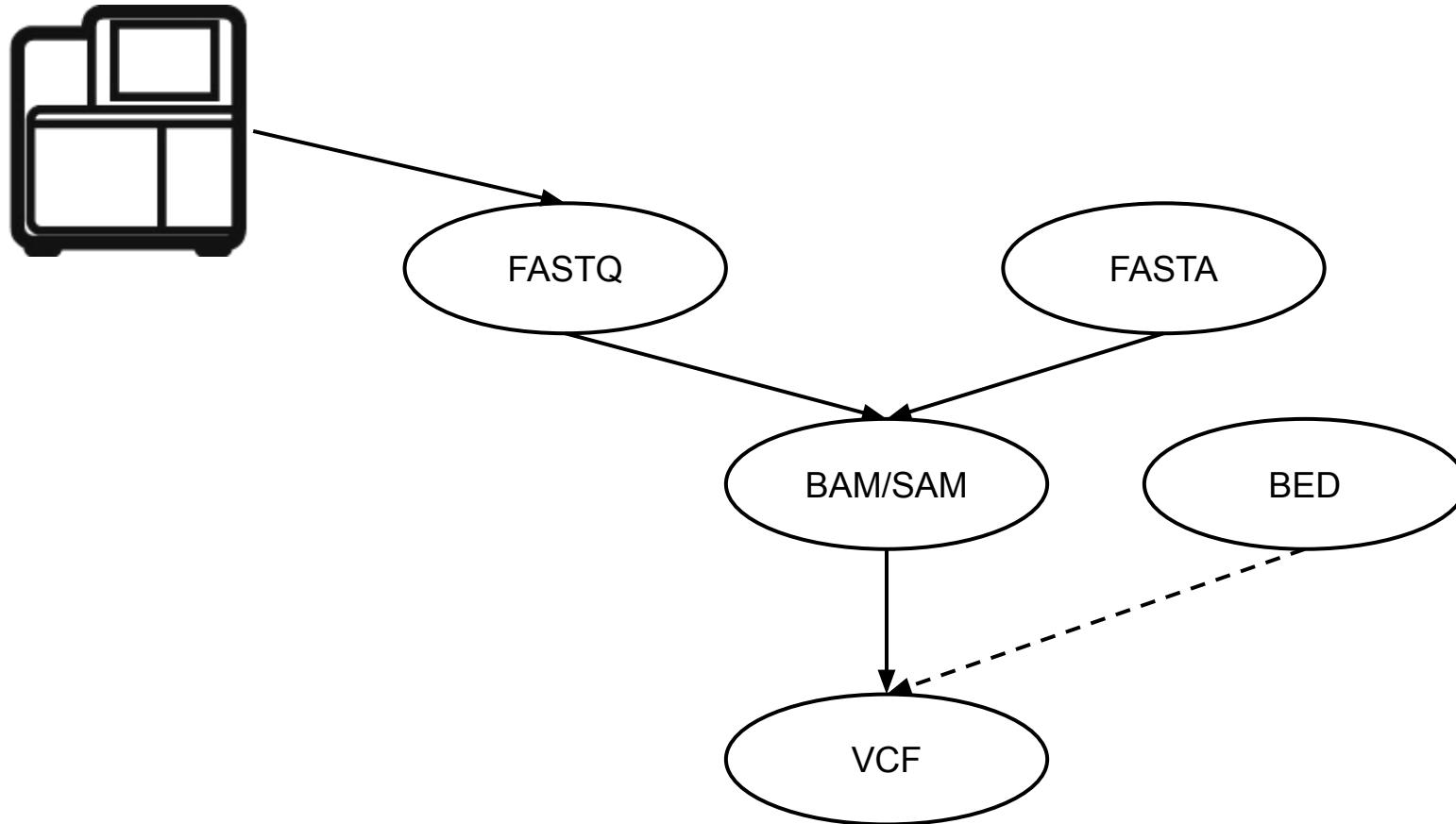


# Analysis workflows

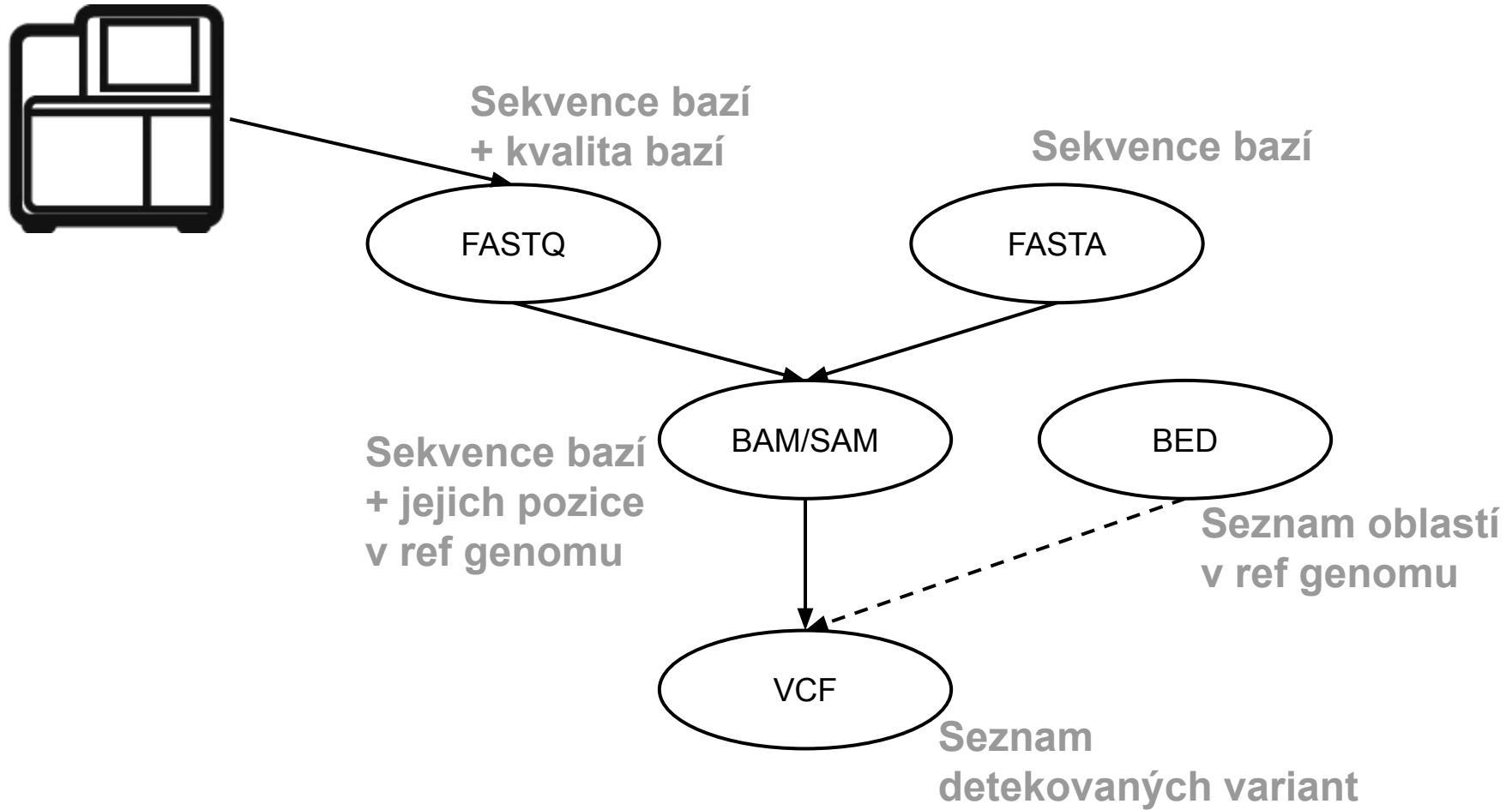


- **anotace variant**
- **efekt SNPs na kódující sekvence**
- **analýza diferenciální exprese**
- **de novo predikce miRNA**
- **identifikace mRNA isoform**
- 
- 
-

# Analysis workflows



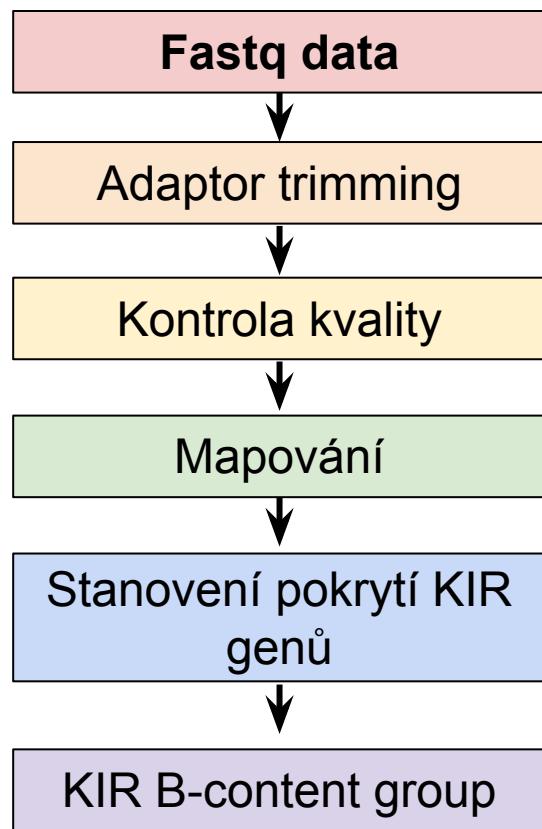
# Analysis workflows



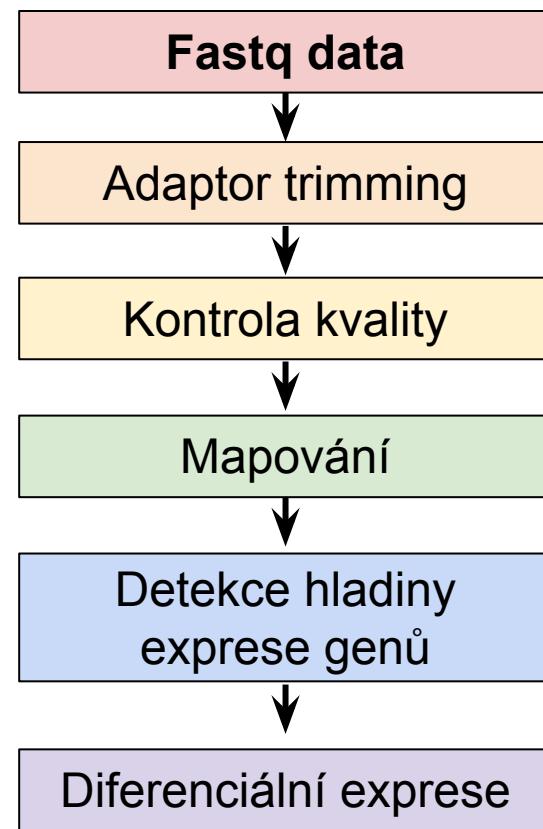
# Program přednášky

- Bioinformatická analýza
- Galaxy
- NGS
- Analysis Workflows
- **Praktická část**
  - **KIR genotyping**
  - **RNAseq analysis**

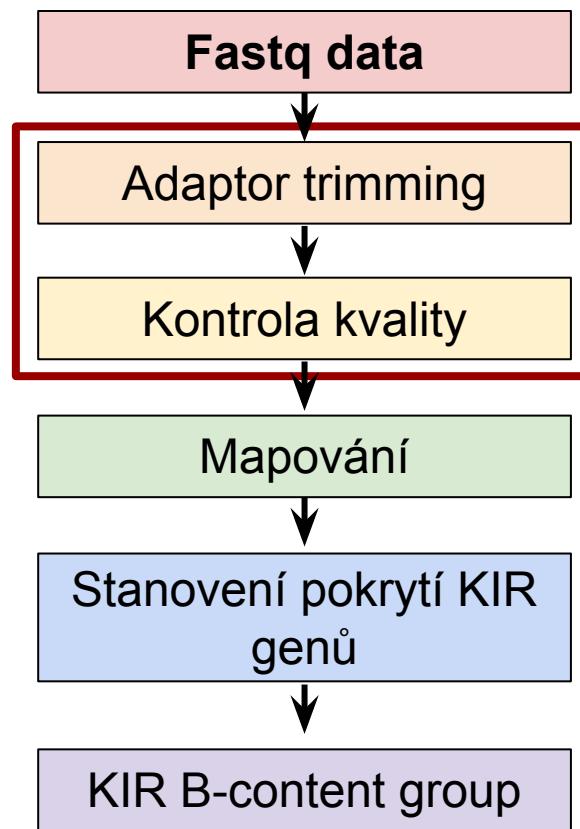
## Genotypizace KIRs



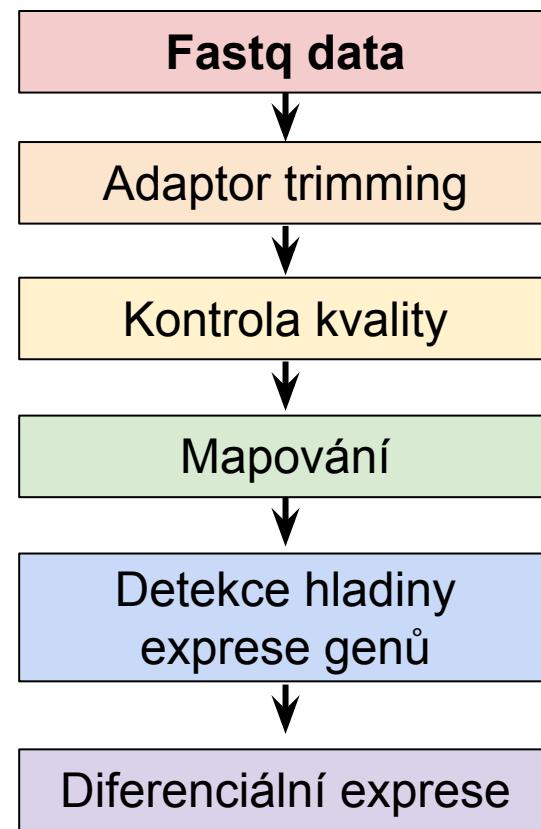
## Diferenciální exprese genů u D. melanogaster



## Genotypizace KIRs



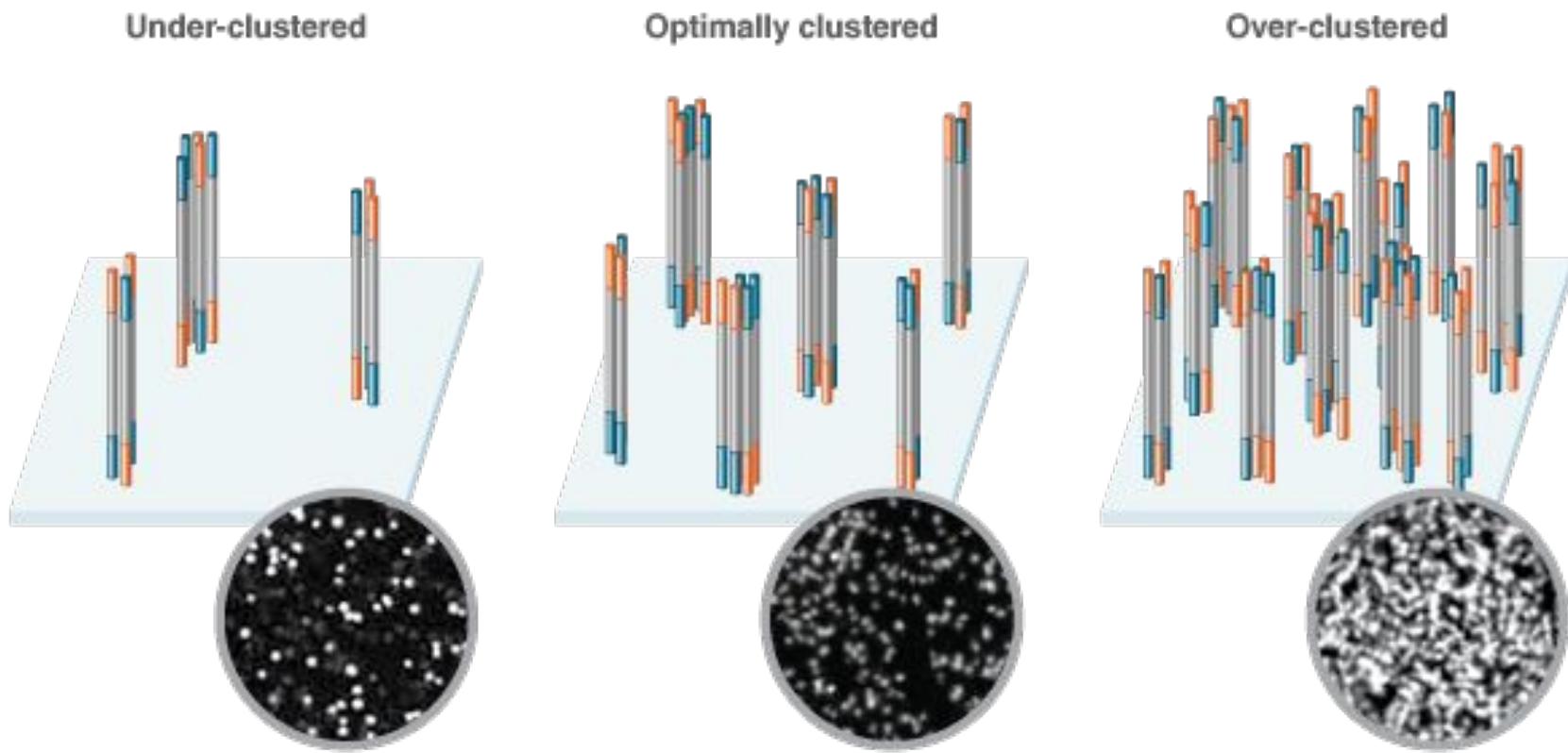
## Diferenciální exprese genů u D. melanogaster



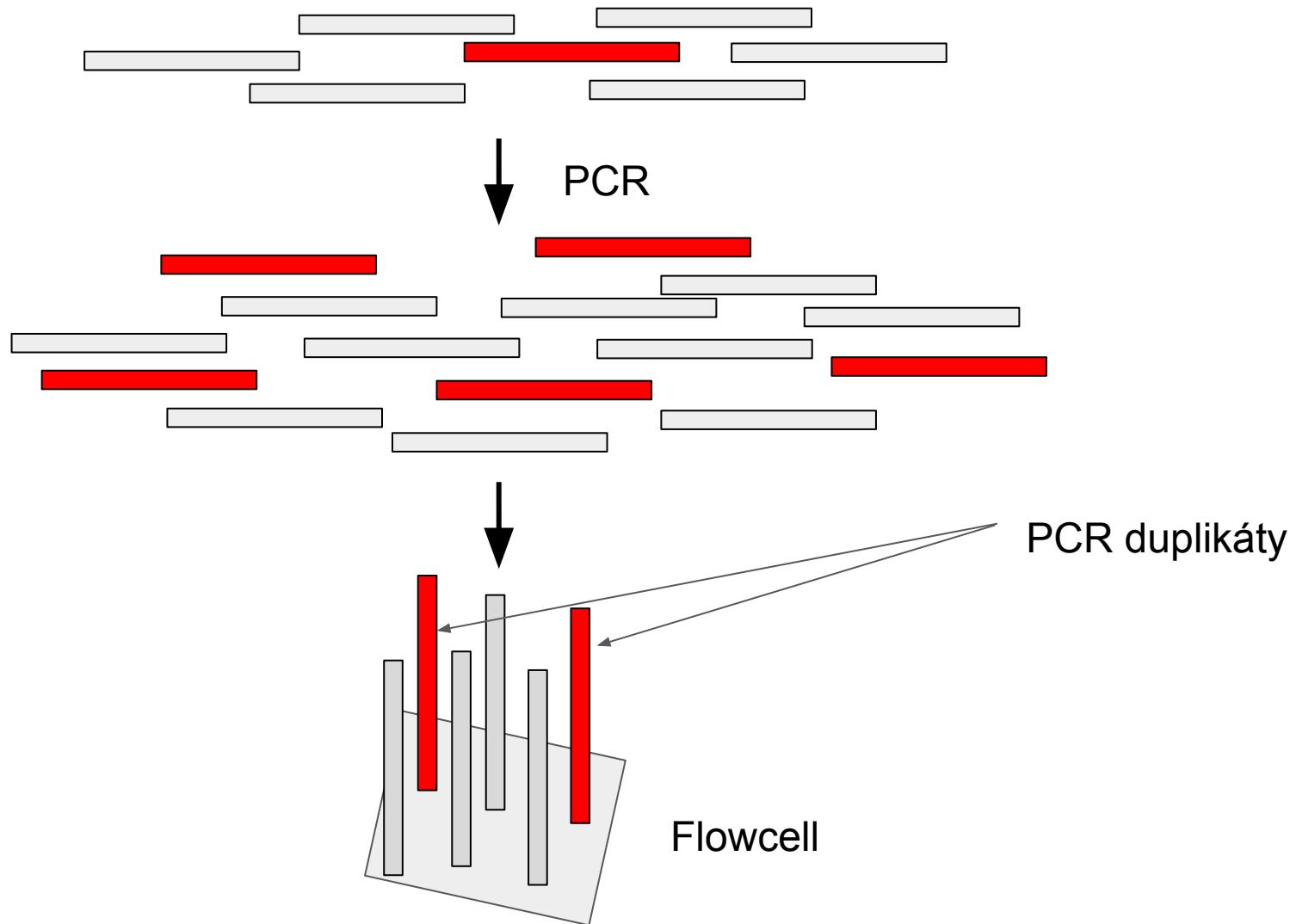
# Illumina seq - co může být špatně?

- Špatné rozpoznání clusterů (overclustering)
- Přítomnost duplikátů
- Kontaminace adaptory
- Asynchronizace molekul clusteru (prephasing a postphasing)
- Plno dalších věcí...

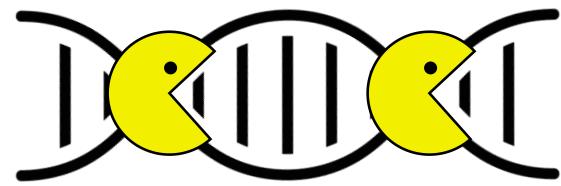
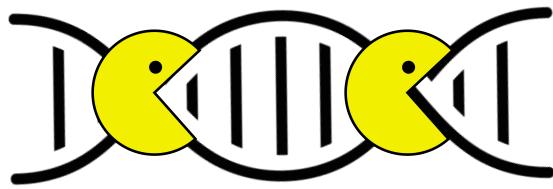
- Špatné rozpoznání svazků (overclustering)



# PCR duplikáty



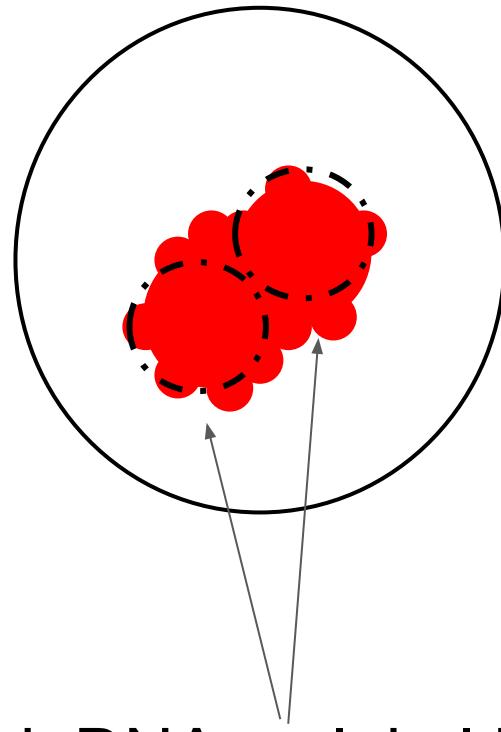
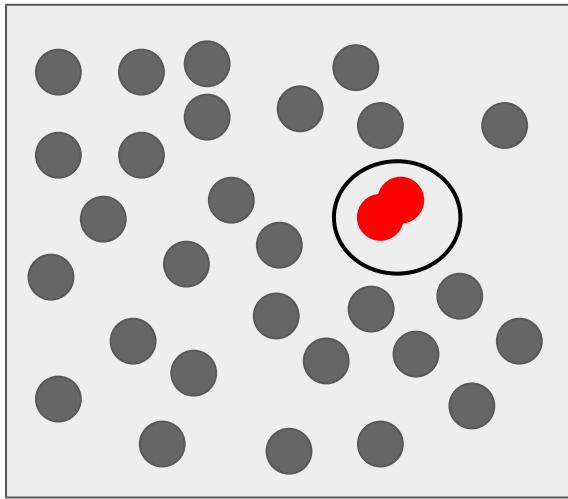
# Štěpné duplikáty



Štěpné duplikáty  
(nenáhodné štěpení)

# Optické duplikáty

Flowcell



1 svazek DNA molekul je brán kamerou jako 2 svazky

# Přítomnost duplikátů

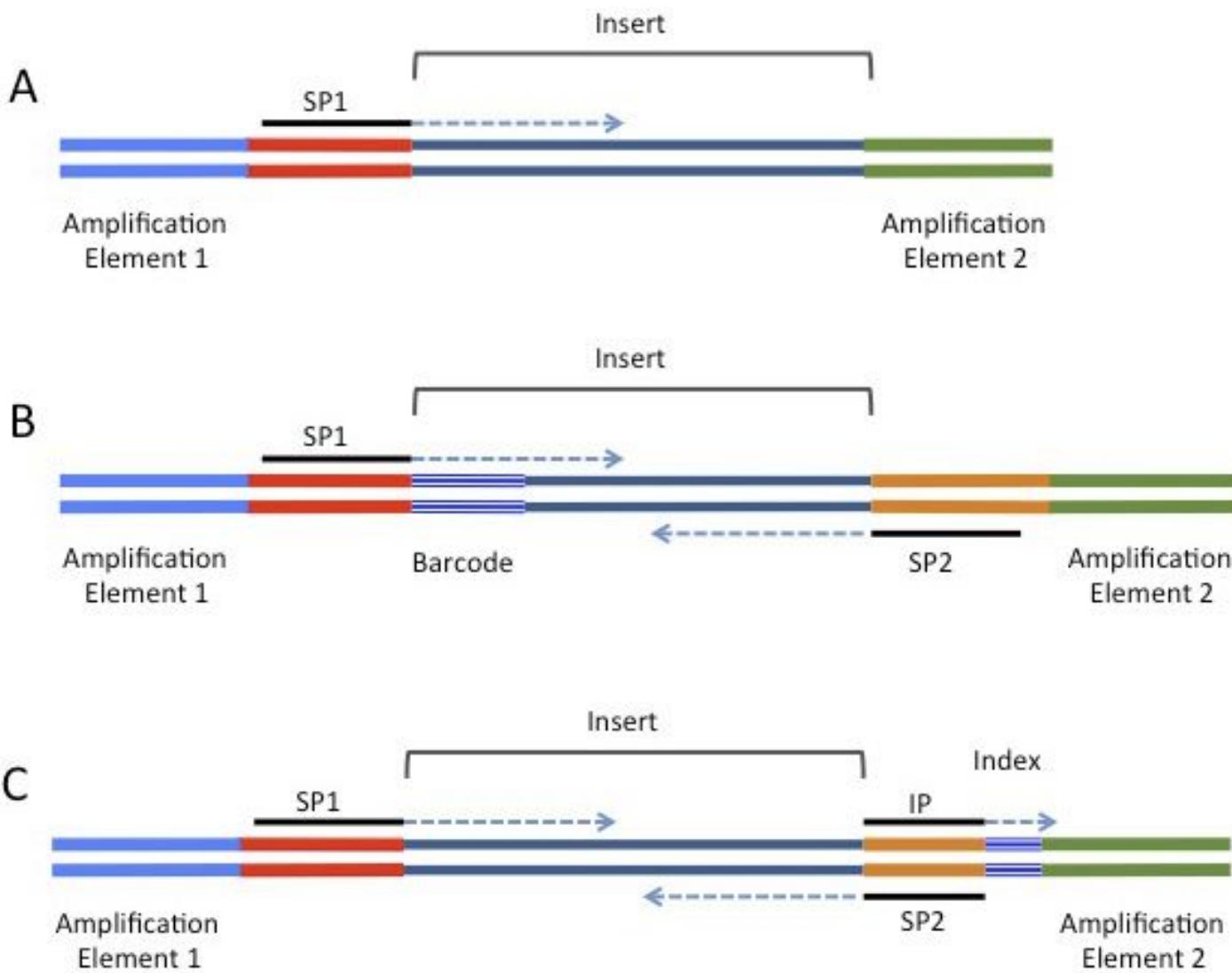
- **PCR duplikáty:** stejné molekuly DNA vzniklé při PCR se dostanou na různá místa flowcell
- **Optické duplikáty:** kamera zaznamená signál ze stejného místa na flowcell několikrát
- **Štěpné duplikáty:** stejné molekuly vznikají během nenáhodného štěpení DNA

# Přítomnost duplikátů

- Duplikáty mohou způsobit zkreslení výsledků, např. změna frekvence alel
- Správně identifikované PCR duplikáty pomáhají opravit chyby PCR a sekvenace

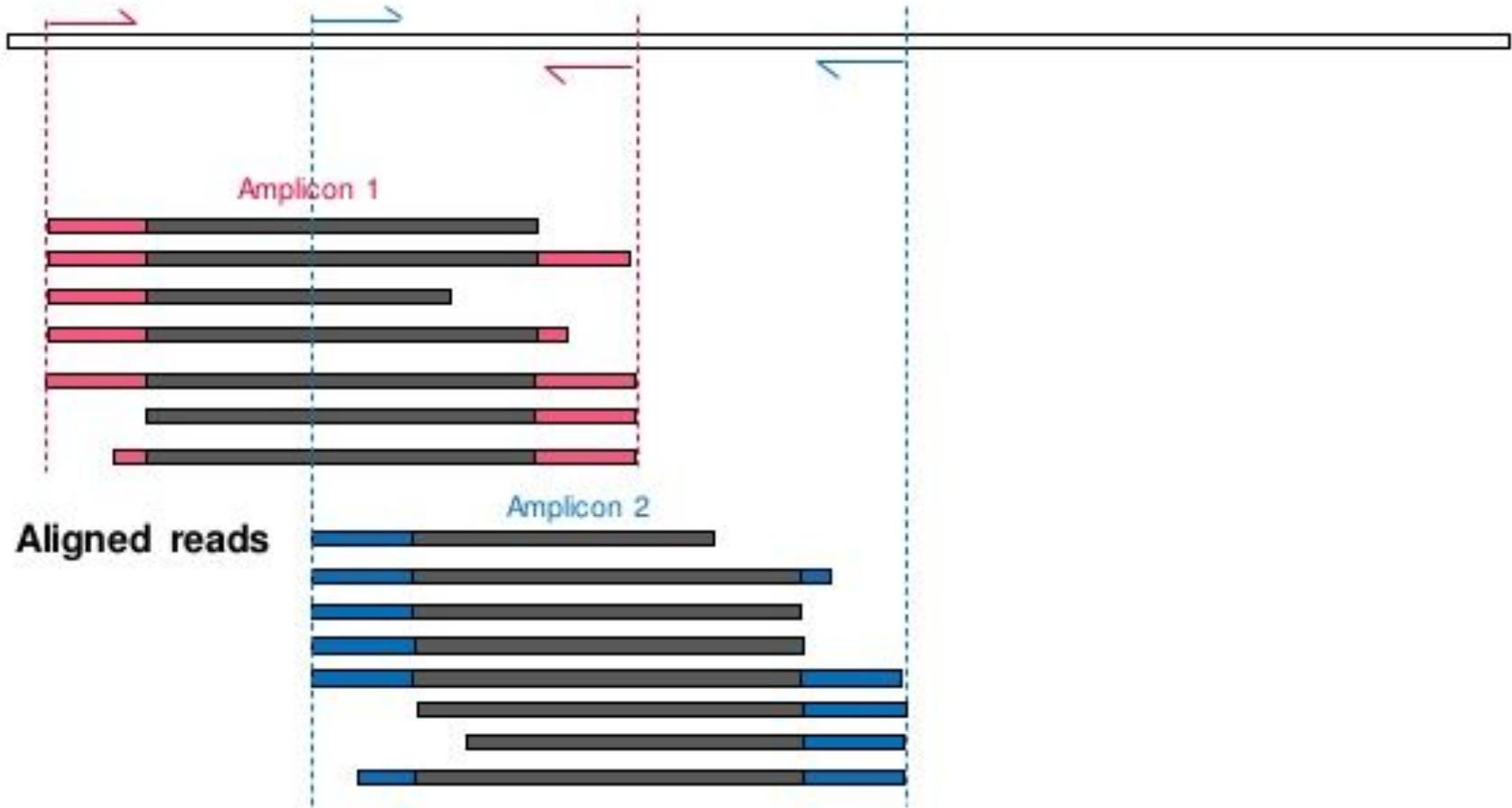


# Kontaminace adaptory



# Primery u amplikonových knihoven

Reference sequence



- Asynchronizace molekul ve svazku  
(prephasing a postphasing)

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

molekuly DNA v  
jednom svazku

Reference

**TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA**

# • Asynchronizace molekul ve svazku (prephasing a postphasing)

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

TGGTT**G**TGATGACTCCTGGCTACTTCCTACTGAA

← Prephasing

Reference

TGGTT**G**TGAT**G**ACTCCT**GG**CTACT**C**CT**A**CT**G**AA

- Asynchronizace molekul ve svazku  
(prephasing a postphasing)

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGBACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA

← Postphasing

Reference

TGGTTGTGATGACTTCGGCTATTCCTAGGAA

# • Asynchronizace molekul ve svazku (prephasing a postphasing)

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA **OK**

TGGTTGTGATGACTCCTGGCTACTTC**C**TACTGAA -3

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA +1

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA **OK**

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA **OK**

TGGTTGTGATGACTCCTGGCTACTT**C**TACTGAA -5

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA +1

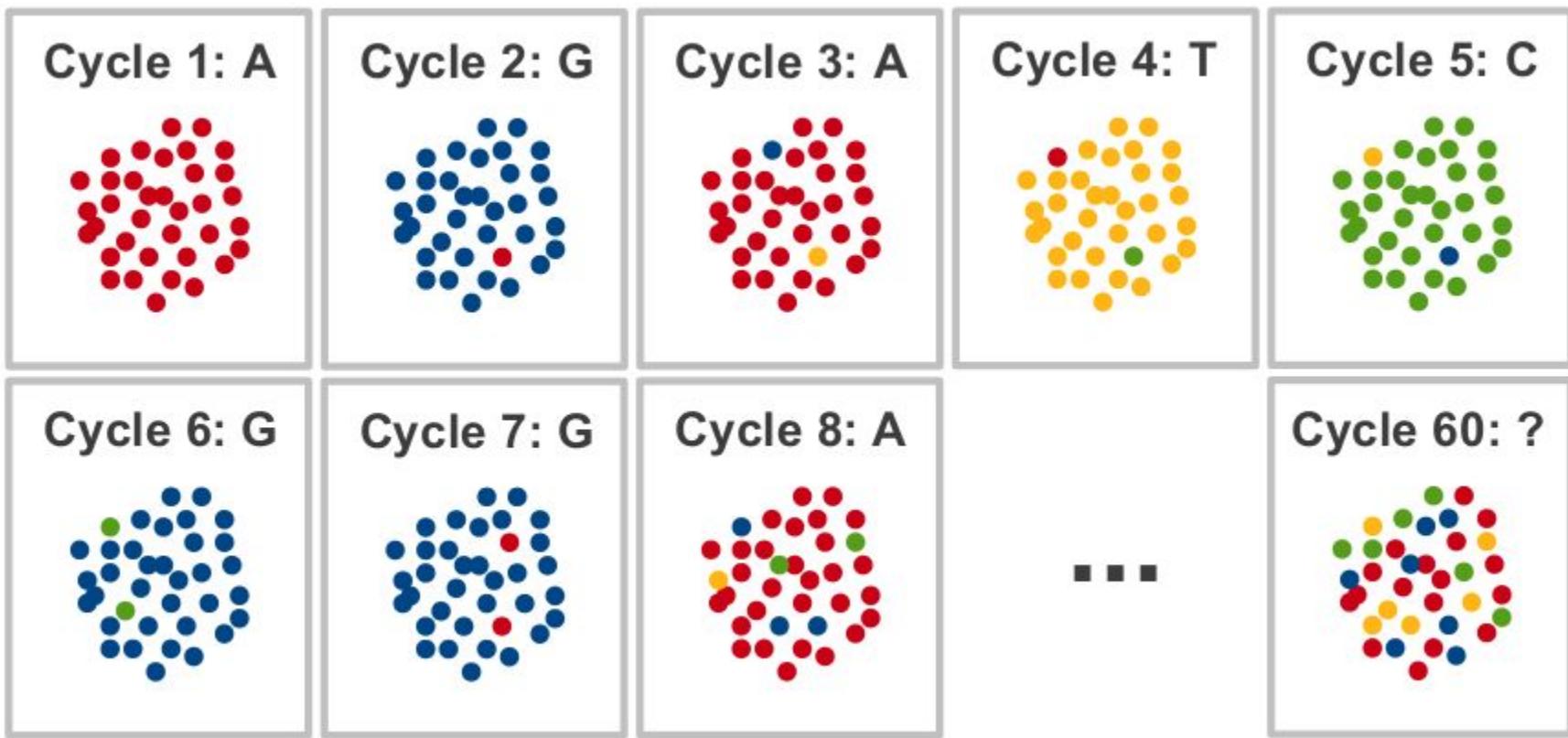
TGGTTGTGATGACTCCTGGCTACTTC**C**TACT**G**A +3

TGGTTGTGATGACTCCTGGCTACTTCCTAC**T**GAA **OK**

Reference

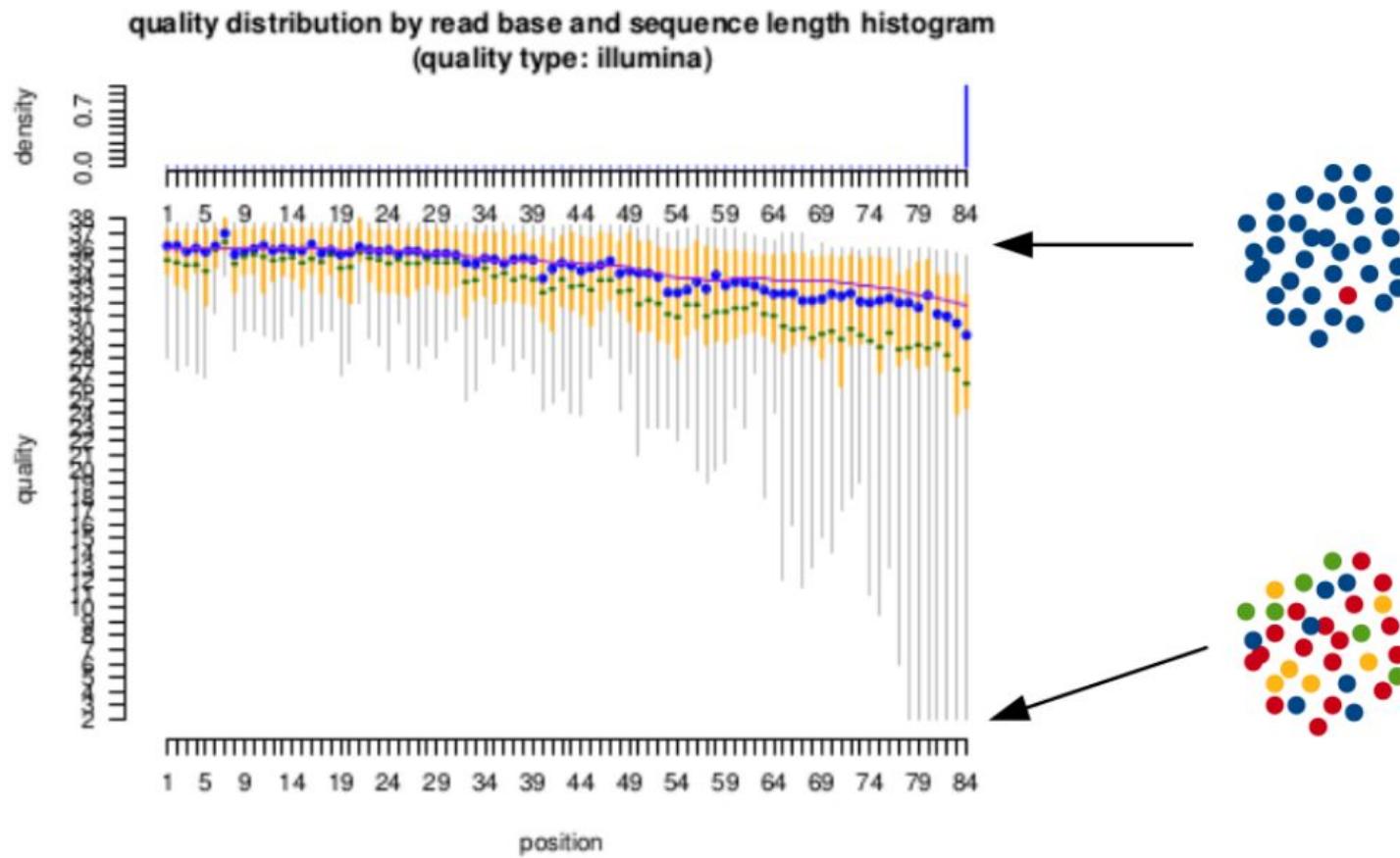
**TGGTTGTGATGACTCCTGGCTACTTCCTACTGAA**

- Asynchronizace molekul ve svazku  
(prephasing a postphasing)



[http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/\\_downloads/Monday\\_JF\\_QAI\\_lecture.pdf](http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf)

- Asynchronizace molekul ve svazku  
(prephasing a postphasing)



[http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/\\_downloads/Monday\\_JF\\_QAI\\_lecture.pdf](http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf)

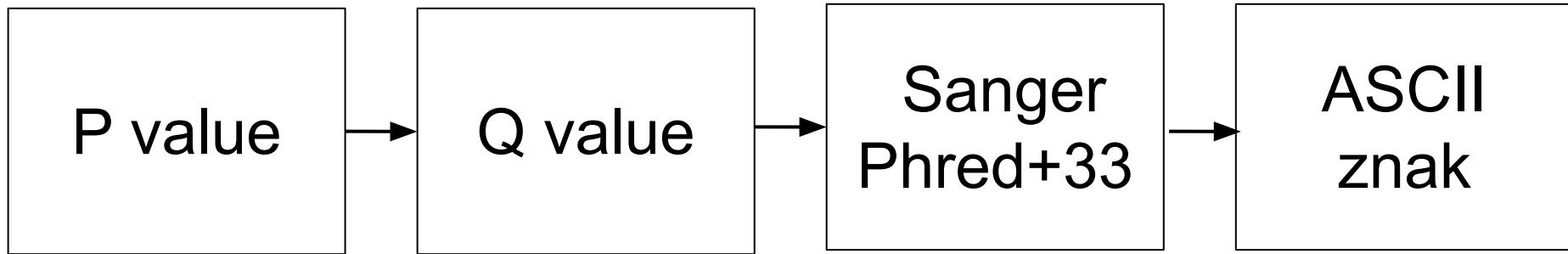
$$Q = -10 \log(P\text{-value})$$

**Phred quality scores are logarithmically linked to error probabilities**

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

# FASTQ formát - kvalita bází

FG**I**EHCCD9DG=1E9?D>CF@HHG??B<GEBGHCG; ;CDB8==C@@>>G**II**@@



0 . 0001

40

73

**I**

# ASCII

Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT End of transmission	0x24	36	\$	0x44	68	D	0x64	100	d
0x05	5	ENQ Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1 Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2 Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3 Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4 Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	FSC Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS File separator	0x3C	60	<	0x5C	92	\	0x7C	124	
0x1D	29	GS Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL



# Různé formáty kvality bází

[https://en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

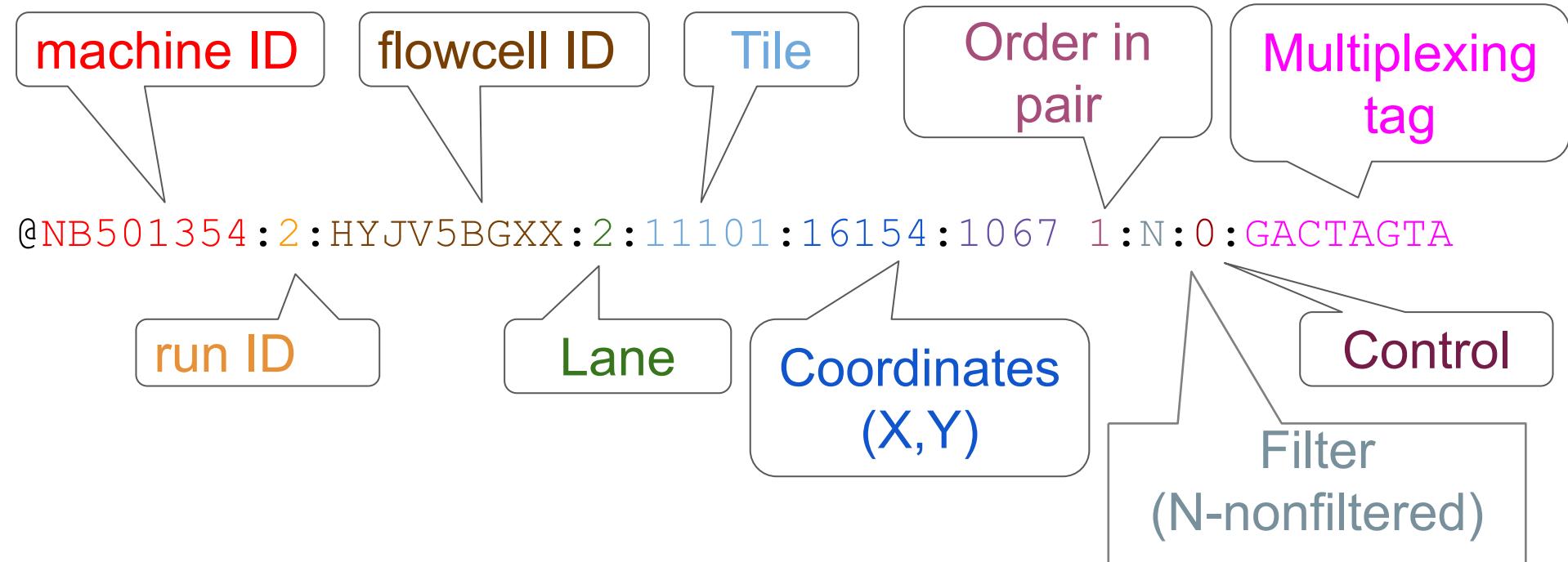
# FASTQ file

Formát pro zápis sekvenčních dat - primární výstup sekvenace

**identifikátor čtení**   **nukleotidová sekvence**   **malé bezvýznamné plus**   **kvalita čtených bazí**

```
@HWI-ST999:102:D1N6AACXX:1:1101:1235:1936 1:N:0:  
ATGTCTCCTGGACCCCTCTGTGCCCAAGCTCCTCATGCATCCTCCTCAGC  
+  
1 : DAADD...<B<AGF=FGIEHCCD9DG=1E9?D>CF@HHG??B<GEBGHC  
@HWI-ST999:102:D1N6AACXX:1:1101:1235:1936 1:N:0:  
ATGTCTCCTGGACCCCTCTGTGCCCAAGCTCCTCATGCATCCTCCTCAGC  
+  
1 : DAADD...<B<AGF=FGIEHCCD9DG=1E9?D>CF@HHG??B<GEBGHC
```

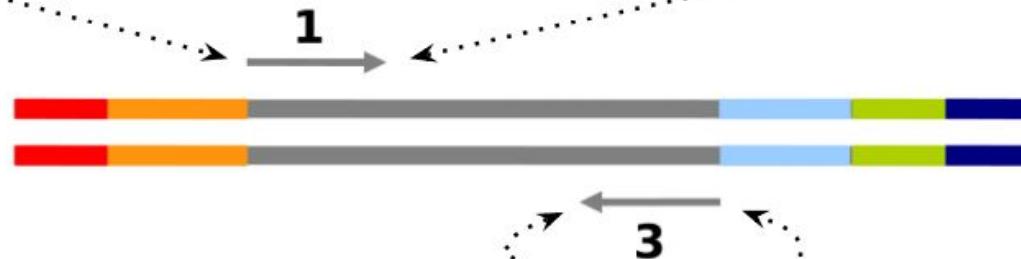
# FASTQ format - read identifier



# Orientace čtení ve FASTQ

```
@HWI-700593F:508:H2FGCBCXX:1:1101:5750:2167 1:N:0:AGTCAA  
AAATGAGATGAAATTGTTCAAGCCAAAGGAGAGCCTTCTTCAACATTCTTCAATTCATCCAT  
+  
BDADBECHHFDECHIHFHFFHIIIIHHFHIHHGHHHIIHHGGHEGGGI1FGGHHHEHH<FCGHEEEFCHICHHCFC
```

An "F/R" pair,  
or "innies"



```
@HWI-700593F:508:H2FGCBCXX:1:1101:5750:2167 2:N:0:AGTCAA  
CATCAAAACTAAAGTTTTAATCCTTACCTTAAACCCCTCAAATTGAAGATCCAAGGCTCCTATCTTGTTCCTTGA  
+  
DADD@HHHFHHHHH1CG?GHFFHHIIIIIIIIHHHHHHHHHIIIEHHHHGGHHIGIIHIGIIIHIIHHHI?H11C
```

[http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/\\_downloads/Monday\\_JF\\_QAI\\_lecture.pdf](http://bioinformatics.ucdavis.edu/docs/2015-september-workshop/_downloads/Monday_JF_QAI_lecture.pdf)

# NGS:QC -> FastQC

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NGS: Gemini

**NGS: QC**

**FastQC Read Quality reports**

NGS: Variant detection

NGS: Alignment

NGS: FASTA nad FASTQ processing

NGS: Variant annotation

NGS: SAMtools

NGS: BAMtools

NGS: BEDtools

NGS: VCF manipulation

Workflows

All workflows

FastQC Read Quality reports (Galaxy Version 0.65)

Short read data from your current history

15: Cutadapt on data 6 and data 5 (Paired Reads)  
14: Cutadapt on data 6 and data 5 (Reads)  
6: H929\_S9\_L001\_R2\_001.fastq  
5: H929\_S9\_L001\_R1\_001.fastq  
4: FASTQ Groomer on data 2

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Execute

**Purpose**

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

**FastQC**

This is a Galaxy wrapper. It merely exposes the external package [FastQC](#) which is documented at [FastQC](#). Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the [Picard-tools](#) libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson. Adaption to version 0.11.2 by T. McGowan.

**Inputs and outputs**

FastQC is the best place to look for documentation - it's very good. A summary follows below for those in a tearing hurry.

This wrapper will accept a Galaxy fastq, sam or bam as the input read file to check. It will also take an optional file containing a list of contaminants information, in the form of a tab-delimited file with 2 columns, name and sequence. As another option the tool takes a custom limits.txt file that allows setting the warning thresholds for the different modules and also specifies which modules to include in the output.

The tool produces a basic text and a HTML output file that contain all of the results, including the following:

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

All except Basic Statistics and Overrepresented sequences are plots.

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

# NGS:QC -> FastQC

Galaxy Using 531.5 MB

Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NGS: Gemini

NGS: QC

NGS: Variant detection

NGS: Alignment

NGS: FASTA nad FASTQ processing

NGS: Variant annotation

NGS: SAMtools

NGS: BAMtools

NGS: BEDtools

NGS: VCF manipulation

Workflows

All workflows

Analyze Data Workflow Shared Data Visualization Admin Help User

H929\_S9\_L001\_R2\_001.fastq FastQC Report

FastQC Report Sat 5 Nov 2016 H929\_S9\_L001\_R2\_001.fastq

History

My first Galaxy analysis 15 shown, 4 deleted

19: FastQC on data 6: Raw Data

18: FastQC on data 6: Web page

383.5 KB format: html, database: hg19

HTML file

17: FastQC on data 5: Raw Data

16: FastQC on data 5: Web page

15: Cutadapt on data 6 and data 5 (Paired Reads)

14: Cutadapt on data 6 and data 5 (Reads)

13: Cutadapt on data 6 and data 5 (Report)

9: UCSC Main on Human: s np138Common (chr22:1-51 304566)

8: UCSC Main on Human: k nownGene (chr22:1-51304 566)

6: H929\_S9\_L001\_R2\_001.f astq

5: H929\_S9\_L001\_R1\_001.f astq

4: FASTQ Groomer on data 2

3: NA18507 sort R2.fastq

2: NA18507 sort R1.fastq

1: 00100-1398177782 Regi ons2.bed

**Summary**

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

**Basic Statistics**

Measure	Value
Filename	H929_S9_L001_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	588540
Sequences flagged as poor quality	0
Sequence length	35-75
%GC	42

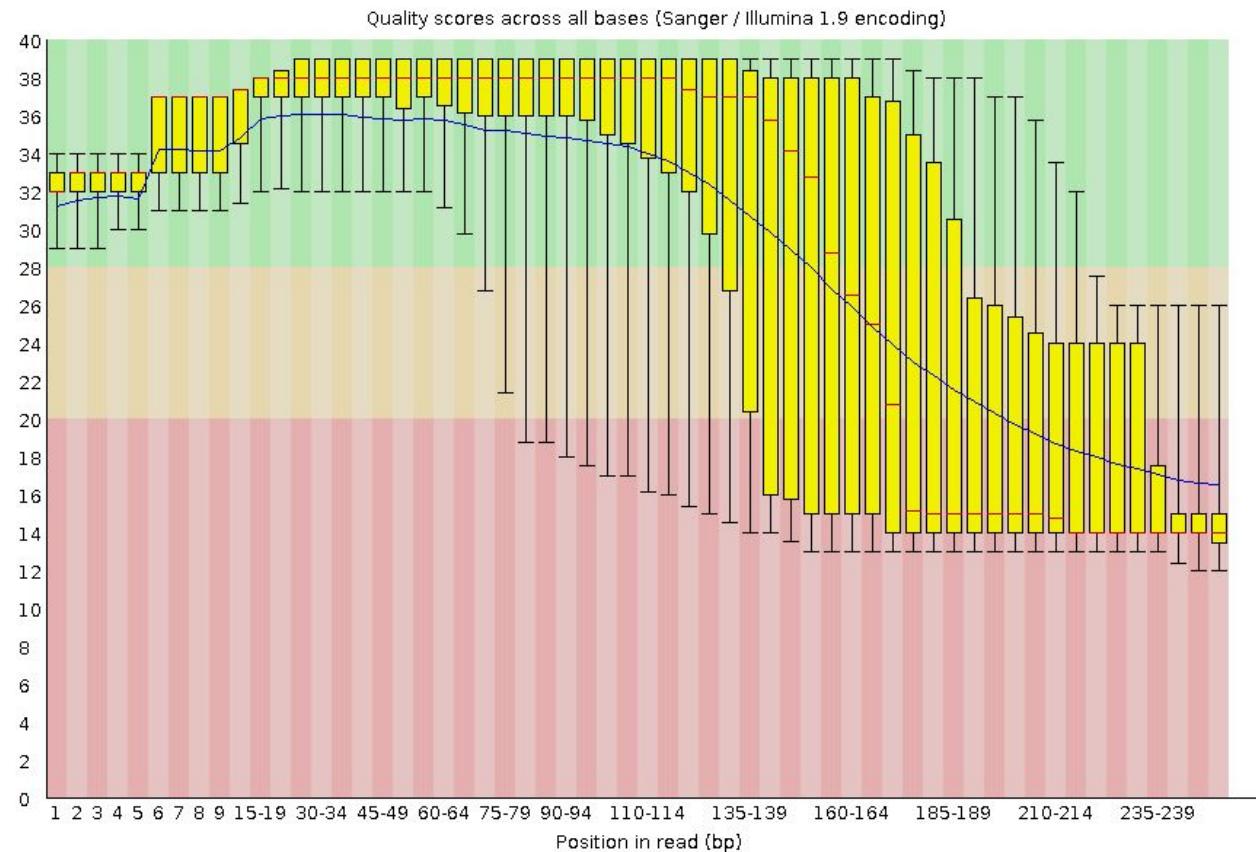
**Per base sequence quality**

Quality scores across all bases (Sanger / Illumina 1.9 encoding)



# Kvalita bází

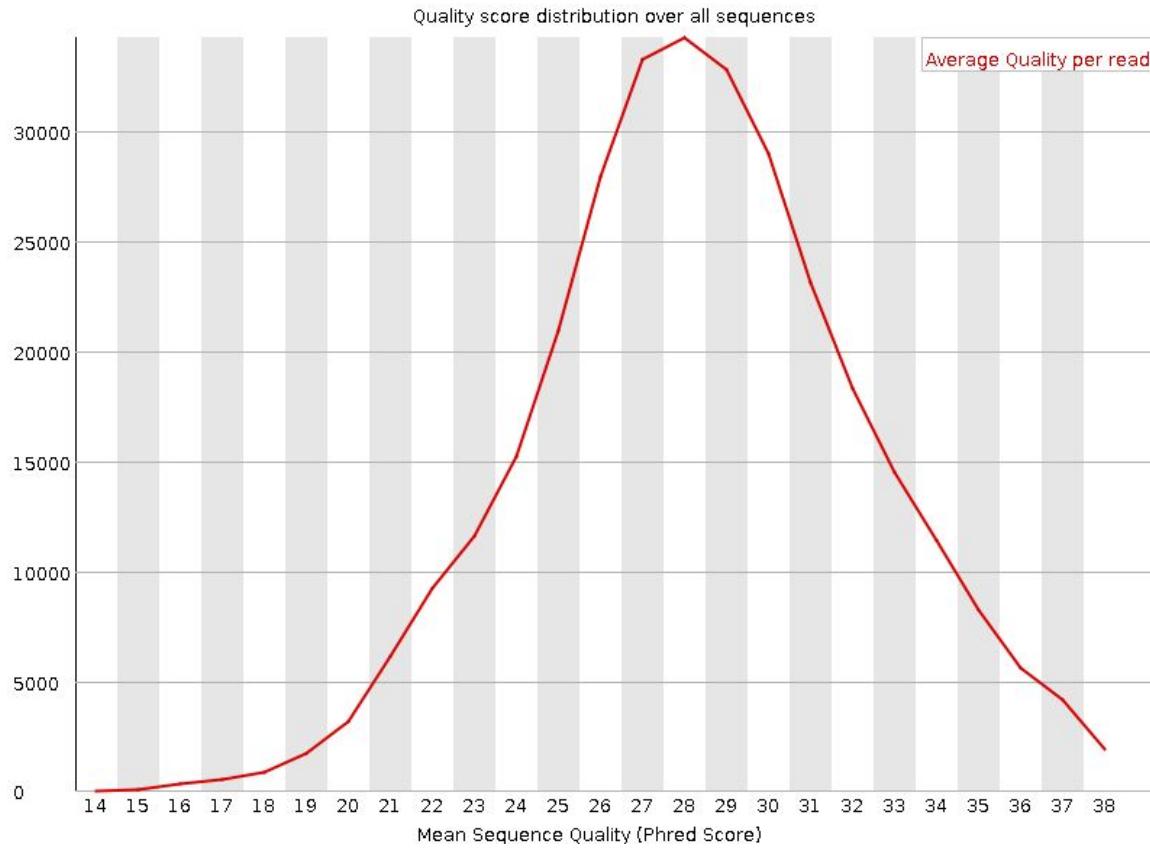
## ✖ Per base sequence quality



# Kvalita čtení



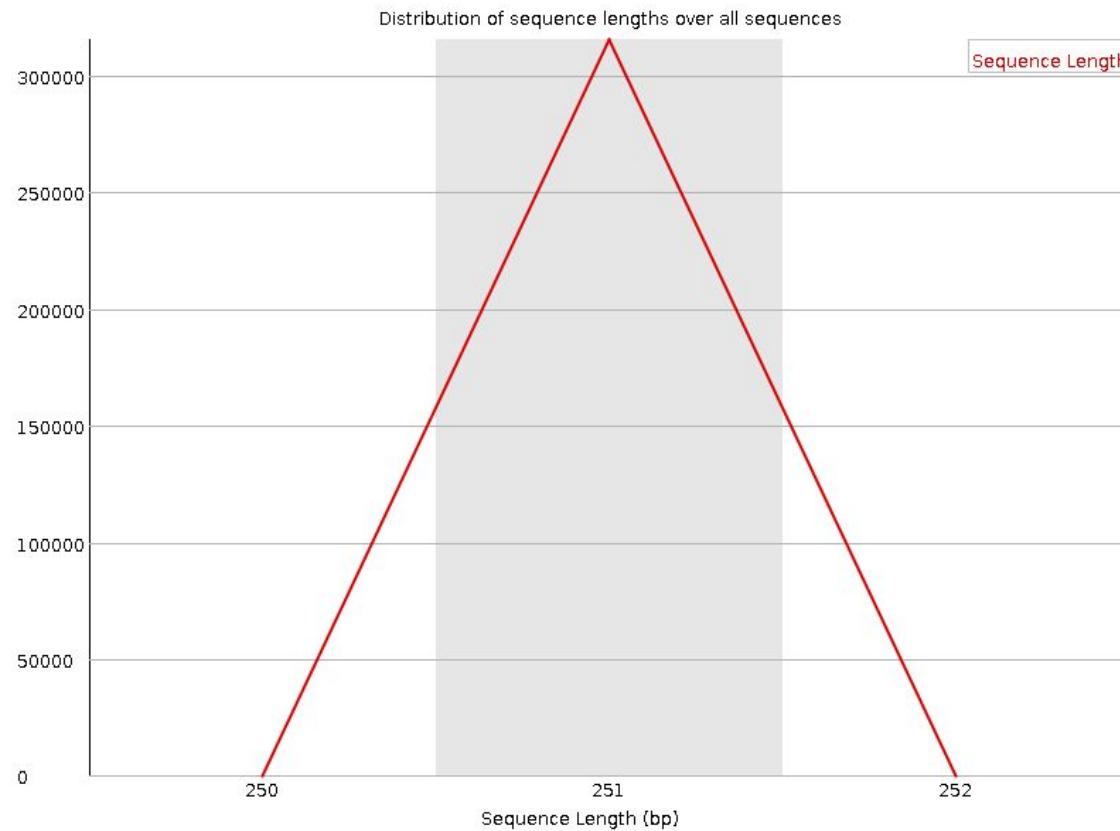
## Per sequence quality scores



# Délka čtení



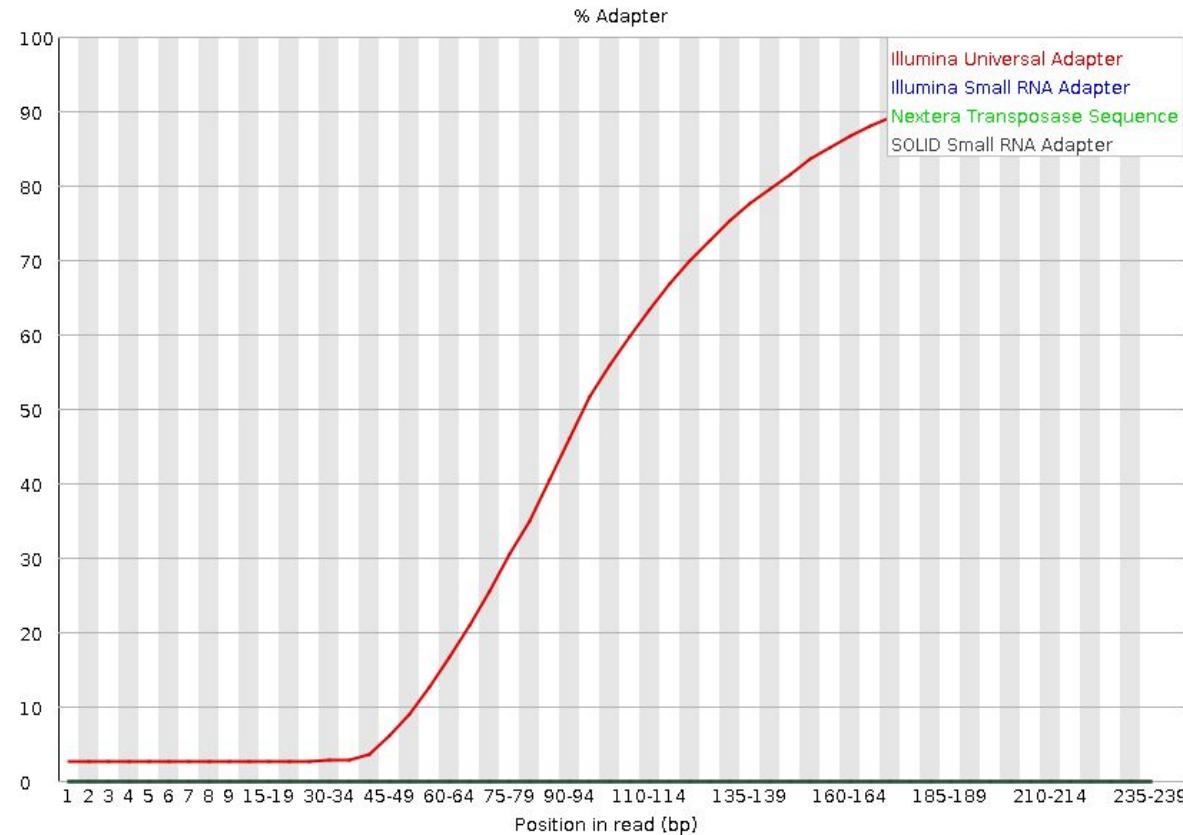
## Sequence Length Distribution



# Kontaminace adaptéry



## Adapter Content



# Ořezání dat - Trimmomatic

Galaxy Using 1.4 GB

Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NGS: Gemini

NGS: QC

NGS: Variant detection

NGS: Alignment

NGS: FASTA nad FASTQ processing

NGS: Variant annotation

NGS: SAMtools

NGS: BAMtools

NGS: BEDtools

NGS: VCF manipulation

Workflows

All workflows

Analyze Data Workflow Shared Data Visualization Admin Help User

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.0)

Paired end data? Yes No

Input Type: Pair of datasets

Input FASTQ file (R1/first of pair): 28: FASTQ Groomer on data 5

Input FASTQ file (R2/second of pair): 29: FASTQ Groomer on data 6

Perform initial ILLUMINACLIP step? Yes No

Cut adapter and other illumina-specific sequences from the read

**Trimmomatic Operation**

1: Trimmomatic Operation

Select Trimmomatic operation to perform: Cut the specified number of bases from the start of the read (HEADCROP)

Number of bases to remove from the start of the read: 11

+ Insert Trimmomatic Operation

Execute

**What it does**

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.

This tool allows the following trimming steps to be performed:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold
- MINLEN: Drop the read if it is below a specified length
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- AVGQUAL: Drop the read if the average quality is below a specified value
- MAXINFO: Trim reads adaptively, balancing read length and error rate to maximise the value of each read

If ILLUMINACLIP is requested then it is always performed first; subsequent options can be mixed and matched and will be performed in the order that they have been specified.

**Inputs**

For single-end data this Trimmomatic tool accepts a single FASTQ file; for paired-end data it will accept either two FASTQ files (R1 and R2), or a dataset collection containing the R1/R2 FASTQ pair.

**Outputs**

For paired-end data a particular strength of Trimmomatic is that it retains the pairing of reads (from R1 and R2) in the filtered output files.

History

My\_first\_Galaxy\_analysis

34 shown, 20 deleted

1.72 GB

54: FastQC on data 43: RawData

53: FastQC on data 43: Webpage

52: Trimmomatic on FA STQ Groomer on data 6

47: FastQC on data 43: RawData

2,408 lines

format: txt database: hg19

46: FastQC on data 43: Webpage

354.8 KB

format: html database: hg19

HTML file

45: Trimmomatic on FA STQ Groomer on data 6 (R2 unpaired)

247 sequences

format: fastqsanger database: hg19

Arguments:

- \* -mx8G
- \* -jar
- \*
- /home/azidkova/galaxy/database/depen 0.36.jar
- \* -PE
- \* -threads
- \* 1
- \* -phred33
- \*
- /home/azidkova/galaxy/database/files/0
- \* /home/azidkova/gala



# Ořezání dat - Trimmomatic

Galaxy Using 1.6 GB

Tools

search tools

Get Data  
Send Data  
Collection Operations  
Text Manipulation  
Filter and Sort  
Join, Subtract and Group  
Convert Formats  
Extract Features  
Fetch Sequences  
Fetch Alignments  
Statistics  
Graph/Display Data  
NGS: Gemini  
NGS: QC  
NGS: Variant detection  
NGS: Alignment  
NGS: FASTA nad FASTQ processing  
NGS: Variant annotation  
NGS: SAMtools  
NGS: BAMtools  
NGS: BEDtools  
NGS: VCF manipulation

Workflows

- All workflows

Analyze Data Workflow Shared Data Visualization Admin Help User

History

search datasets

My\_first\_Galaxy\_analysis 34 shown, 24 deleted 1.92 GB

58: Trimmomatic on FA STQ Groomer on data 6 (R2 unpaired)

57: Trimmomatic on FA STQ Groomer on data 5 (R1 unpaired)

56: Trimmomatic on FA STQ Groomer on data 6 (R2 paired)

55: Trimmomatic on FASTQ Groomer on data 5 (R1 paired)

54: FastQC on data 43: RawData

53: FastQC on data 43: Webpage

52: Trimmomatic on FA STQ Groomer on data 6

51: FastQC on data 35: RawData

50: FastQC on data 35: Webpage

49: FastQC on data 35: RawData

48: FastQC on data 35: Webpage

47: FastQC on data 43: RawData

46: FastQC on data 43: Webpage

45: FastQC on data 43: RawData

44: FastQC on data 43: Webpage

43: FastQC on data 43: RawData

42: FastQC on data 43: Webpage

41: FastQC on data 35: RawData

40: FastQC on data 35: Webpage

39: FastQC on data 43: RawData

38: FastQC on data 43: Webpage

37: FastQC on data 43: RawData

36: FastQC on data 43: Webpage

35: FastQC on data 43: RawData

34: FastQC on data 43: Webpage

33: FastQC on data 43: RawData

32: FastQC on data 43: Webpage

31: FastQC on data 43: RawData

30: FastQC on data 43: Webpage

29: FastQC on data 43: RawData

28: FastQC on data 43: Webpage

27: FastQC on data 43: RawData

26: FastQC on data 43: Webpage

25: FastQC on data 43: RawData

24: FastQC on data 43: Webpage

23: FastQC on data 43: RawData

22: FastQC on data 43: Webpage

21: FastQC on data 43: RawData

20: FastQC on data 43: Webpage

19: FastQC on data 43: RawData

18: FastQC on data 43: Webpage

17: FastQC on data 43: RawData

16: FastQC on data 43: Webpage

15: FastQC on data 43: RawData

14: FastQC on data 43: Webpage

13: FastQC on data 43: RawData

12: FastQC on data 43: Webpage

11: FastQC on data 43: RawData

10: FastQC on data 43: Webpage

9: FastQC on data 43: RawData

8: FastQC on data 43: Webpage

7: FastQC on data 43: RawData

6: FastQC on data 43: Webpage

5: FastQC on data 43: RawData

4: FastQC on data 43: Webpage

3: FastQC on data 43: RawData

2: FastQC on data 43: Webpage

1: FastQC on data 43: RawData

0: FastQC on data 43: Webpage

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.



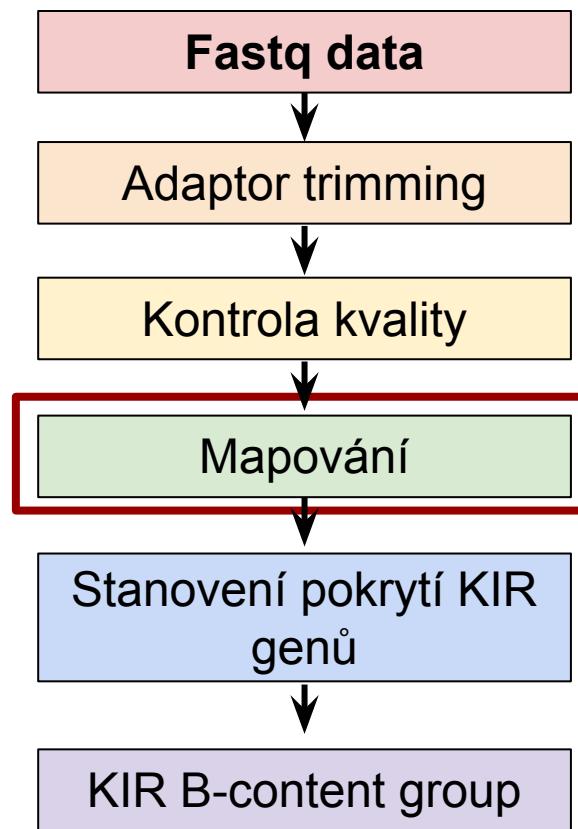
# Cvičení 1 - kontrola kvality a trimming

1. V Galaxy importujte historii “**KIR-QC**” ([https://usegalaxy.org/u/tomas\\_hron/h/kir-qc](https://usegalaxy.org/u/tomas_hron/h/kir-qc)).
2. Spusťte kontrolu kvality FastQ souborů pomocí nástroje “**FastQC**”. **Byl u dat proveden “adaptor trimming”?**
3. Filtrujte data pomocí “**Trimmomatic**” (IlluminaClip, TruSeq3 PE adaptory;TRAILING, kvalita>20).
4. Proveďte kontrolu kvality na ořezaných datech dle bodu 3. **Jaká je minimální délka čtení? Kolik procent párových čtení nám zbylo z původního počtu?**

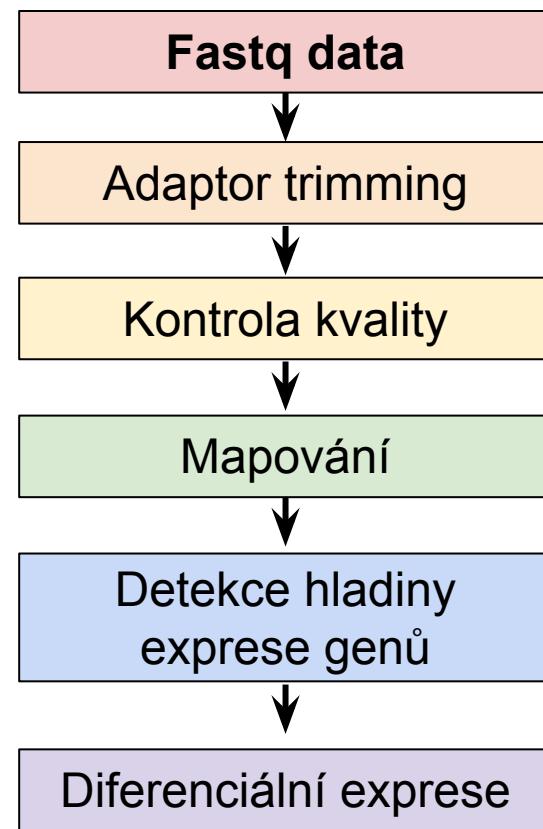
\*[https://usegalaxy.org/u/tomas\\_hron/h/kir-qc-finished](https://usegalaxy.org/u/tomas_hron/h/kir-qc-finished)

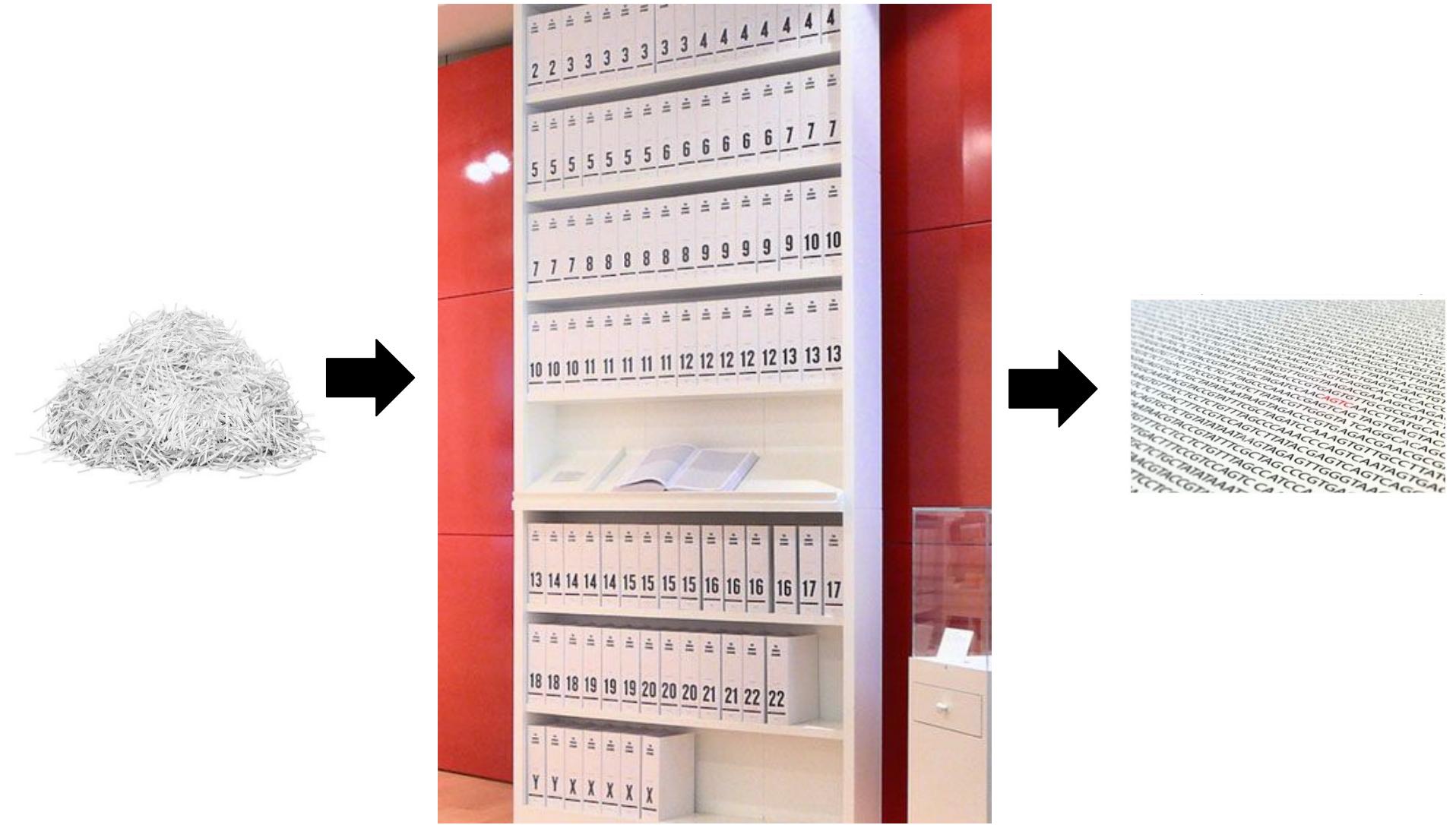


## Genotypizace KIRs



## Diferenciální exprese genů u D. melanogaster





The first printout of the human genome to be presented as a series of books, displayed at the [Wellcome Collection](#), London

# Genome reference sequence

<b>Release name</b>	<b>Date of release</b>	<b>Equivalent UCSC version</b>
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

# Genome reference sequence

**GRCh37.p13:**

Total bases:

*3.23 Billion*

*2.99 Billion (without N)*

N50:

*46 Million*

Number of alternative loci:

*9*

Non-nuclear genome:

*No*

**GRCh38.p2:**

Total bases:

*3.21 Billion*

*3.05 Billion (without N)*

N50:

*67 Million*

Number of alternative loci :

*261*

Non-nuclear genome:

*Yes*

# Reference transcriptome sequences

## RefSeq

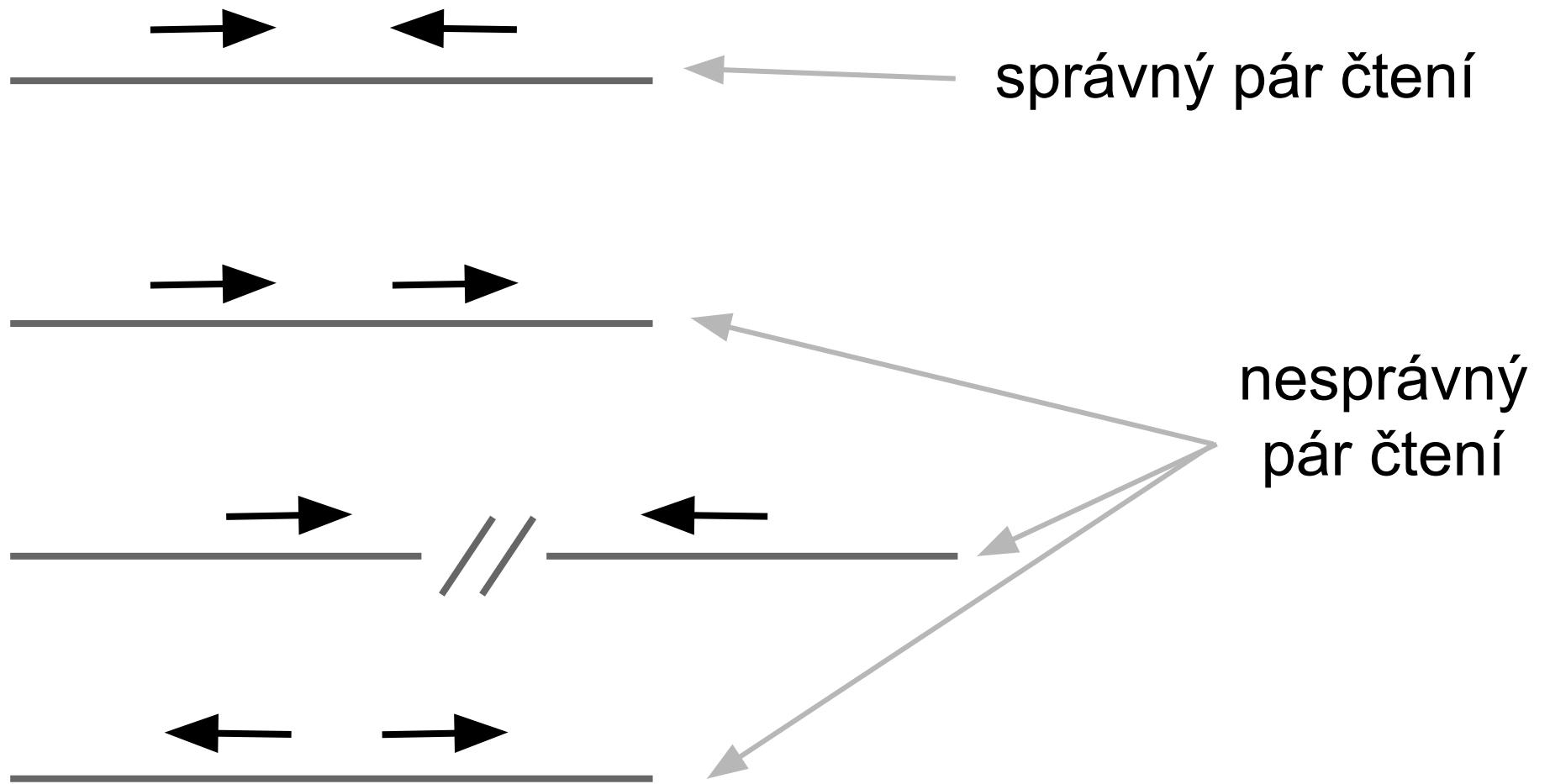
The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially [RefSeqGene](#) records), expression studies, and comparative analyses.

## GENCODE

For human and mouse, this combined Ensembl/HAVANA gene set is the default gene set from the GENCODE project. For both human and mouse, we also guarantee that all transcripts from the Consensus Coding Sequence (CCDS) set are present in the GENCODE gene set. The complete GENCODE gene set for human and mouse can be viewed on our website in the GENCODE Comprehensive track. A subset of these transcript models are tagged as being in the GENCODE Basic set, which is also available on our website.



# Párová čtení (paired-end reads)



# Edit distance

AACGT-AGCC Reference  
A---GTCA<sup>T</sup>CC Čtení

Edit distance = 4

# Mapping quality

The calculation of mapping qualities considers all the factors below:

- The repeat structure of the reference. Reads falling in repetitive regions usually get very low mapping quality.
- The base quality of the read. Low quality means the observed read sequence is possibly wrong, and wrong sequence may lead to a wrong alignment.
- Paired end or not. Reads mapped in pairs are more likely to be correct.



# Mapping quality

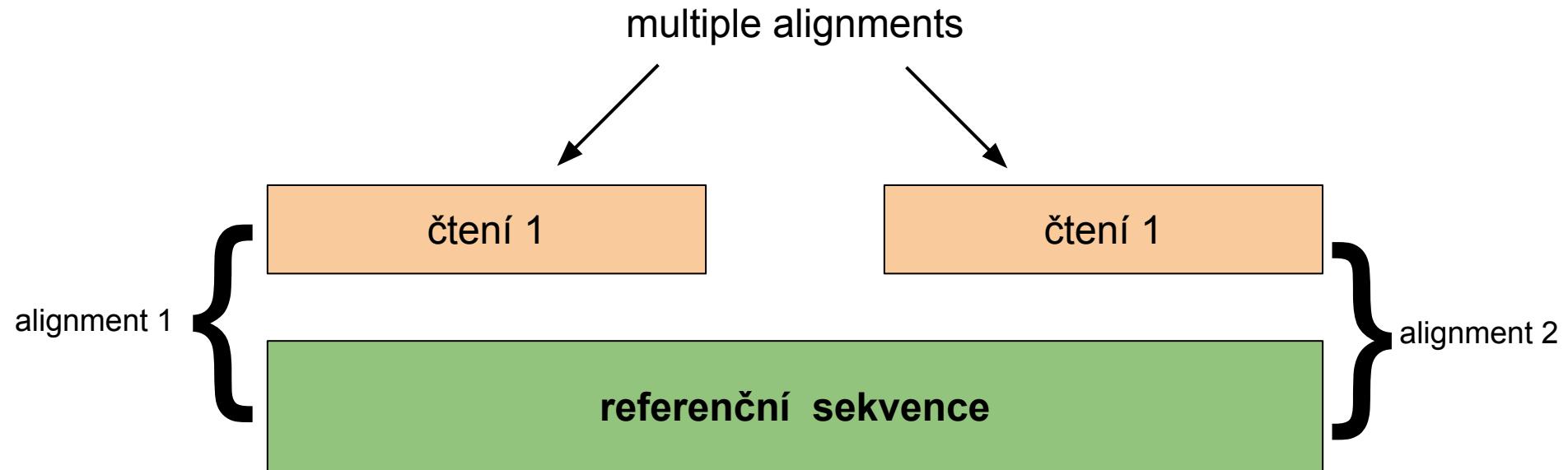
Reference ..... C C C G C C G G A A A T T .....

Read C C G C C G G G A A

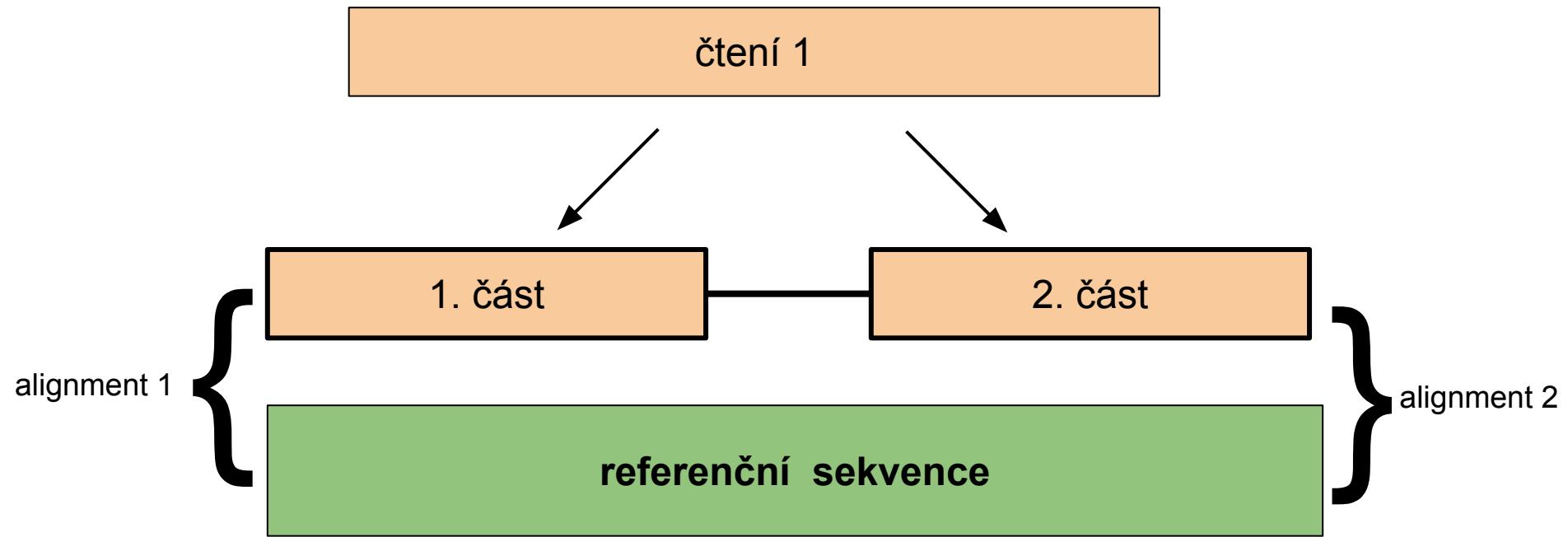
Reference C C C G C C G G A A A T T .....

Alignments (1) C C G C C G G G A A MQ=40 (3) C C G C C G G G A A MQ=50

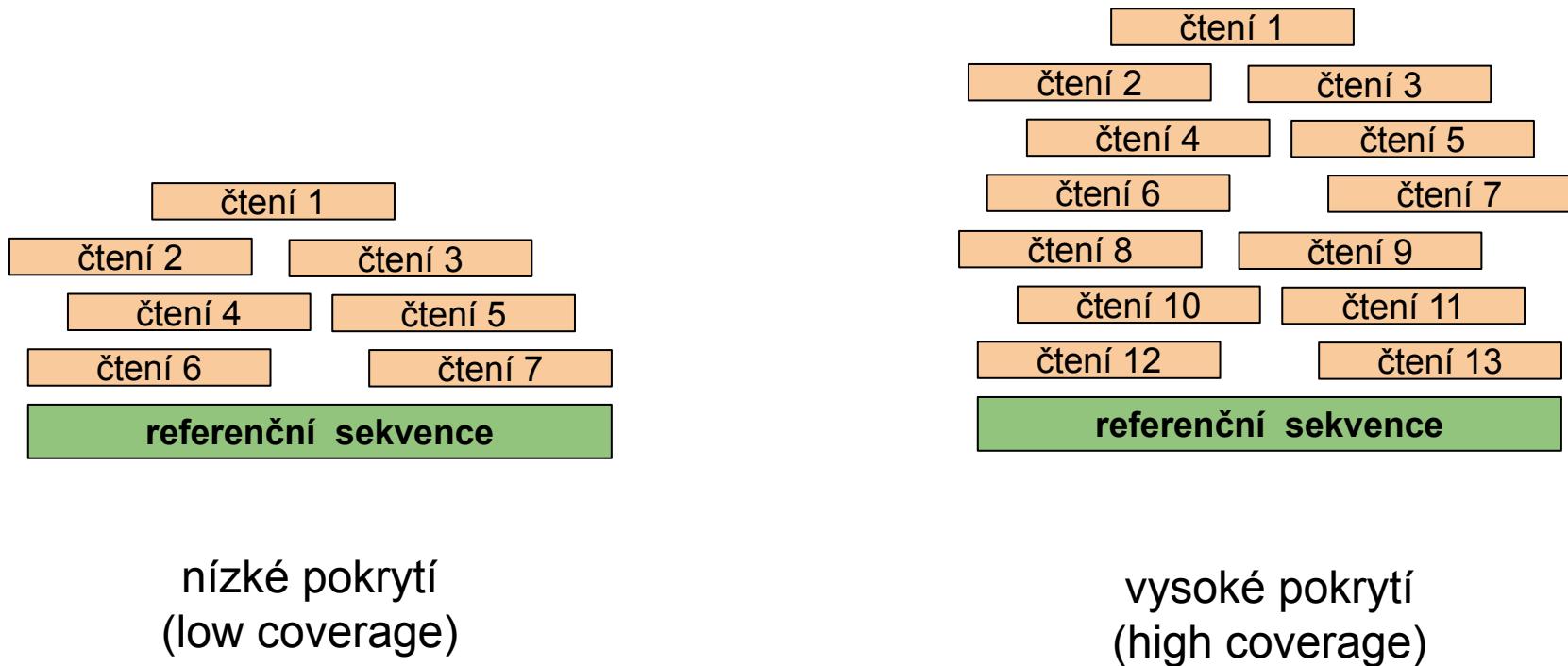
# secondary alignment



# split alignment



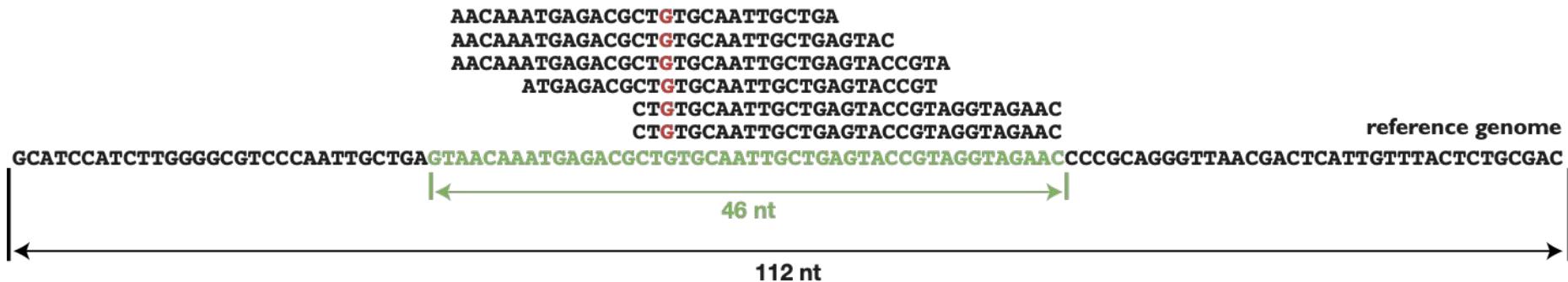
# Hloubka sekvenačního pokrytí



nízké pokrytí  
(low coverage)

vysoké pokrytí  
(high coverage)

# Hloubka sekvenačního pokrytí



Reference point	Calculation	Example (see Fig.2)
Whole genome	(# of sequenced bases)* / (genome size)	188 / 112 = 1,68 fold
One locus	(# of bases mapping to the locus) / (size of locus)	188 / 46 = 4,09 fold
One position	(# of reads overlapping with one position)	6 fold

# SAM/BAM (Sequence/Binary Alignment Map)

- Hlavička (řádky začínají s @)
  - HD - header
  - SQ - reference sequence dictionary
  - RG - read group
  - PG - program
- Tabulka (1 read alignment na každý řádek)
  - QNAME - název čtení
  - FLAG - informace ohledně mapování
  - RNAME - název referenční sekvence
  - POS - pozice, na kterou bylo namapované čtení (nejvíce vlevo)
  - MAPQ - kvalita namapování čtení na referenční sekvenci (phred)
  - CIGAR - zapis alignmentu
  - RNEXT - název párového čtení (pokud stejny tak “=”, pokud není - “\*”)
  - PNEXT - pozice párového čtení (pokud není tak “0”)
  - TLEN - délka templátu / sekv. fragmentu (pokud není tak “0”)
  - SEQ - sekvence čtení
  - QUAL - kvalita čtení
  - TAGS - nepovinná doplňující data
- Více informací: <https://samtools.github.io/hts-specs/SAMv1.pdf>



# SAM - příklad

hlavička

```
@HD VN:1.0 SO:coordinate
@SQ SN:1 LN:249250621 AS:NCBI37
UR:file:/data/GATK/human_g1k_v37.fasta
@RG ID:UM0098:1 PL:ILLUMINA PU:HWUSI-L001 LB:80 ...
@PG ID:bwa VN:0.5.4
```

alignment 1

```
1:497:R:-272+13M17D24M 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;=====9;>>>>=>>>>>>>=>>>>>>>> XT:A:U NM:i:0 SM:i:37 AM:i:0
X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
```

alignment 2

```
19:20389:F:275+18M2D19M 99 1 17644 0 37M = 17919 314
TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>><>><>>4::>>:<9 RG:Z:UM0098:1 XT:A:R NM:i:0
SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37
```



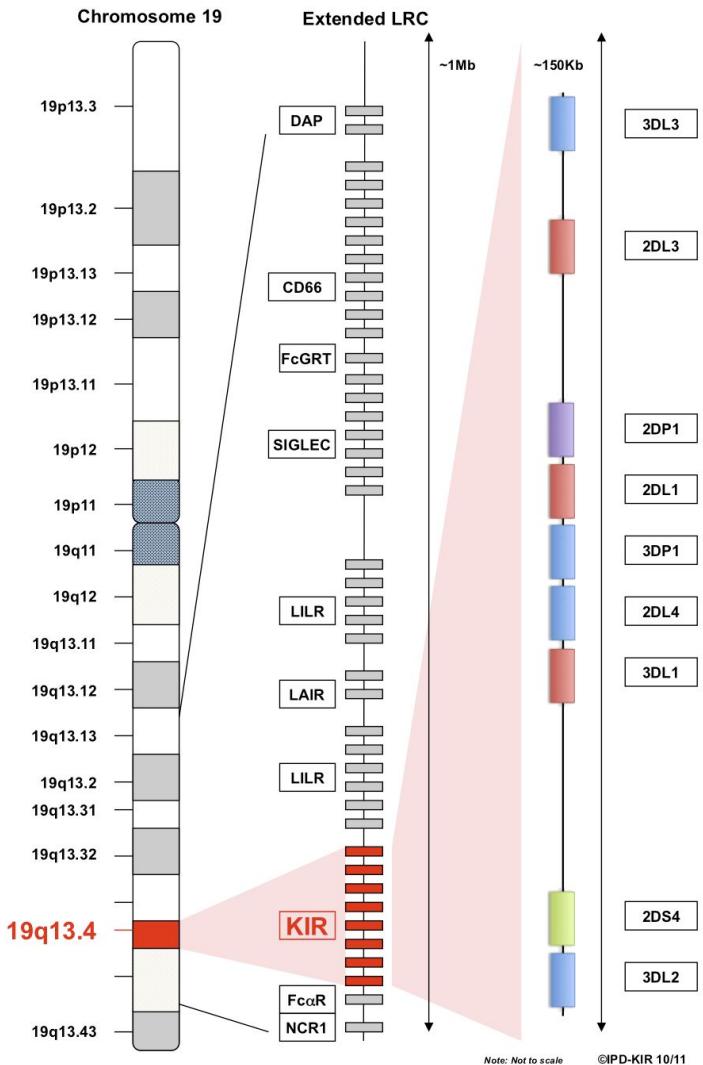
HPST, s.r.o.



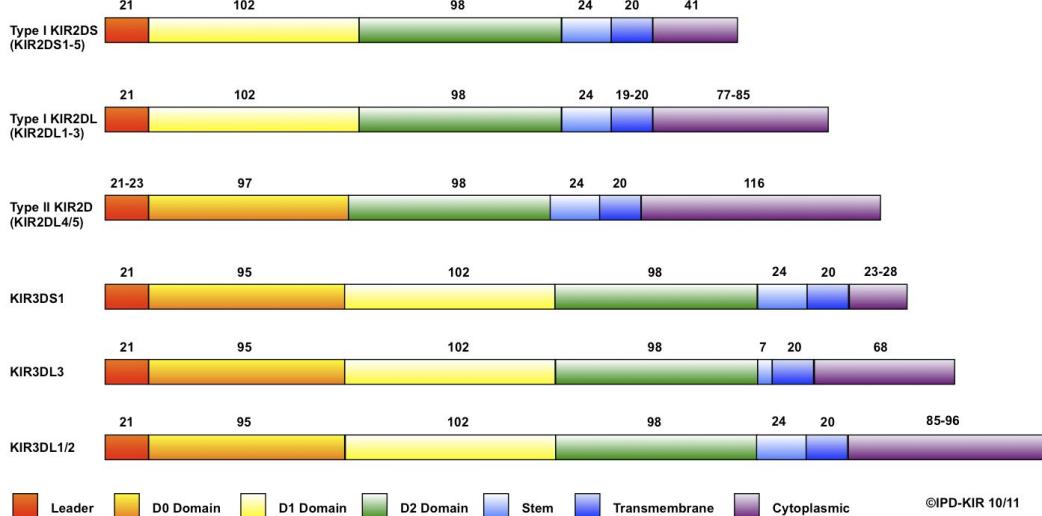
# Úskalí mapování NGS dat

- Chybovost sekvenování
  - Illumina ~0.1%
- Variabilita
  - individual genome differs from the reference human genome at 4.1 million to 5.0 million sites affecting 20 million bases (1000 Genomes Project)
- Repetitivní oblasti
  - ~50% genomu

# Jakou referenci použít pro KIRs?



The KIR gene family currently consists of 15 gene loci (KIR2DL1, KIR2DL2/L3, KIR2DL4, KIR2DL5A, KIR2DL5B, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, KIR3DL1/S1, KIR3DL2, KIR3DL3 and two pseudogenes, KIR2DP1 and KIR3DP1) encoded within a 100-200 Kb region of the Leukocyte Receptor Complex (LRC) located on chromosome 19 (19q13.4) (1).



# Jakou referenci použít pro KIRs?

EMBL-EBI

Services

Research

Training

About us



## IPD-KIR

Overview

IMGT/HLA

KIR

MHC

NHKIR

HPA

ESTDAB

Contact

Support

IPD / KIR

## IPD-KIR

### Release 2.9.0, 11 December 2019

The database provides a centralised repository for human KIR sequences. Killer-cell Immunoglobulin-like Receptors (KIR) have been shown to be highly polymorphic at the allelic and haplotypic level. KIRs are members of the immunoglobulin superfamily (IgSF) formerly called Killer-cell Inhibitory Receptors. They are composed of two or three Ig-domains, a transmembrane region and cytoplasmic tail which can in turn be short (activatory) or long (inhibitory). The Leukocyte Receptor Complex (LRC) which encodes KIR genes has been shown to be polymorphic, polygenic and complex like the MHC.



HPST, s.r.o.

Agilent Technologies  
Autorizovaný distributor

# Jakou referenci použít pro KIRs?

## Resources

- [!\[\]\(1ad53017baba922c95a6c4c0e7fbd21c\_img.jpg\) About >](#)
- [!\[\]\(1b82a9d16c0a94acede90e029ee28e0f\_img.jpg\) Statistics >](#)
- [!\[\]\(dfb753276996173d13fb95b507cc7171\_img.jpg\) Publications >](#)
- [!\[\]\(9406d437e2a8e698fd2ed2c7e70b9a08\_img.jpg\) Nomenclature >](#)
- [!\[\]\(b1c53fdea0da4e1370e49fdb7a23145d\_img.jpg\) Genes >](#)
- [!\[\]\(caa0645b6b46fce406dabc4fdf10ee75\_img.jpg\) Alleles >](#)
- [!\[\]\(d421e4f7870933dd88a7d7f36c1393af\_img.jpg\) Haplotypes >](#)
- [!\[\]\(787b0041f4fd678acaf44535a037ff82\_img.jpg\) Genotypes >](#)
- [!\[\]\(9f86cd0295c718e5bb9870356a8276fa\_img.jpg\) Alignments >](#)
- [!\[\]\(86a196b7381a5ff2fc48a1ee6e9aaafe\_img.jpg\) Releases >](#)
- [!\[\]\(f0248f07d73a6f3fe45c3835214a934e\_img.jpg\) BLAST >](#)
- [!\[\]\(cea000c74a0bb1670c1214a7c0b08bbc\_img.jpg\) B-Content >](#)
- [!\[\]\(3ab00c18284632fda9e739c95e479fe6\_img.jpg\) Ligands >](#)
- [!\[\]\(841575275c4110efb7023d91057712a5\_img.jpg\) FTP >](#)
- [!\[\]\(5a8f085dd40c43bbc4a4de11f156cf82\_img.jpg\) GitRepos >](#)
- [!\[\]\(e7b0512d7d5f72bbd7cc55a27282dc7e\_img.jpg\) Cells >](#)
- [!\[\]\(8857a35eca2487b875a40a5af3429c6e\_img.jpg\) FAQ >](#)
- [!\[\]\(aca02cf80bbb8244ea1d9aa87be9f3d0\_img.jpg\) Submissions >](#)



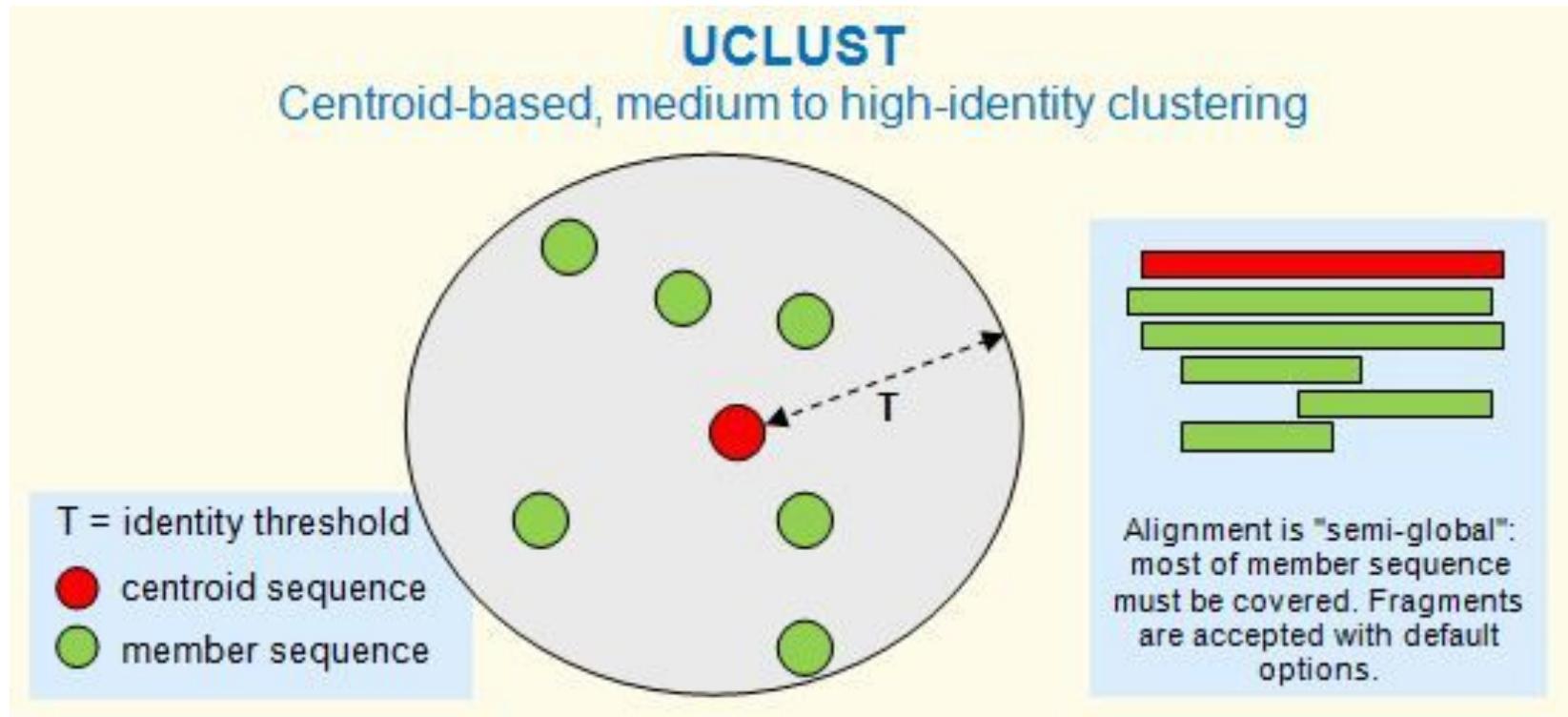
# Jakou referenci použít pro KIRs?

## Resources

- [!\[\]\(b110472768e4beffc0a22de70a03d082\_img.jpg\) About >](#)
- [!\[\]\(8a015d8e36be8a03d65d50d56f929111\_img.jpg\) Publications >](#)
- [!\[\]\(db2b117a89a17ee250c8178ec7afc1c2\_img.jpg\) Genes >](#)
- [!\[\]\(228330ffbe42ccbd30292910be3244b5\_img.jpg\) Haplotypes >](#)
- [!\[\]\(b897670275f554ca98a68b9750bb4e8f\_img.jpg\) Alignments >](#)
- [!\[\]\(582ea77737fff1f420913569b4715963\_img.jpg\) BLAST >](#)
- [!\[\]\(5649487d21c3d73fcf1c76db4cd47a41\_img.jpg\) Ligands >](#)
- [!\[\]\(d45ad1a72adb8a356500c0bf2b86d410\_img.jpg\) GitRepos >](#)
- [!\[\]\(fb80bb18a95087efb55a481cd859f585\_img.jpg\) FAQ >](#)
- [!\[\]\(a311821b2afc8b100374729b1ec34e71\_img.jpg\) Statistics >](#)
- [!\[\]\(29fcaf75776ac5464c145b69654009e5\_img.jpg\) Nomenclature >](#)
- [!\[\]\(d8424f5359880bba58cf4c7c9b8c8de2\_img.jpg\) Alleles >](#)
- [!\[\]\(c3a3c664804689513225e6d9490baa2b\_img.jpg\) Genotypes >](#)
- [!\[\]\(a950cb3521ea71b76d1cad37775709fb\_img.jpg\) Releases >](#)
- [!\[\]\(ef818b77ed8df008411f9dbb2ad15f5d\_img.jpg\) B-Content >](#)
- [!\[\]\(88e45de1fe256578edf5be7bb107c1ac\_img.jpg\) FTP >](#)
- [!\[\]\(0dd8906653892541d506cdffb074770e\_img.jpg\) Cells >](#)
- [!\[\]\(ac9940b01d781a0315f8535564c821c5\_img.jpg\) Submissions >](#)

# Jakou referenci použít pro KIRs?

## Homology-based clustering



# Jakou referenci použít pro KIRs?

## Homology-based clustering

VSearch clustering (Galaxy Version 2.8.3.0)

Favorite Versions Options

Select your input FASTA file

6: VSearch clustering on data 2: Cluster centroids

Choose sorting method to use before clustering

Cluster sequences after sorting by length (--cluster-fast)

Indicate that input sequences are not presorted by length

Yes No

(--usersort)

ID definition

int

(--iddef)

Reject hit if identity is lower than this value



HPST, s.r.o.



# Cvičení\_2 - clustering DNA sekvencí KIR receptorů

1. Importujte historii “**KIR-Clustering**” ([https://usegalaxy.org/u/tomas\\_hron/h/kir-clustering](https://usegalaxy.org/u/tomas_hron/h/kir-clustering)).
2. Proveďte Clustering u obou souborů zvlášť pomocí “**VSearch clustering**”:
  - Změňte verzi nástroje na Galaxy Version 1.9.7.0
  - Reject hit if identity is lower than this value: 0.95
  - Mask sequences: No masking
  - Write cluster abundances to centroid file. Yes
  - Select output files: Centroids

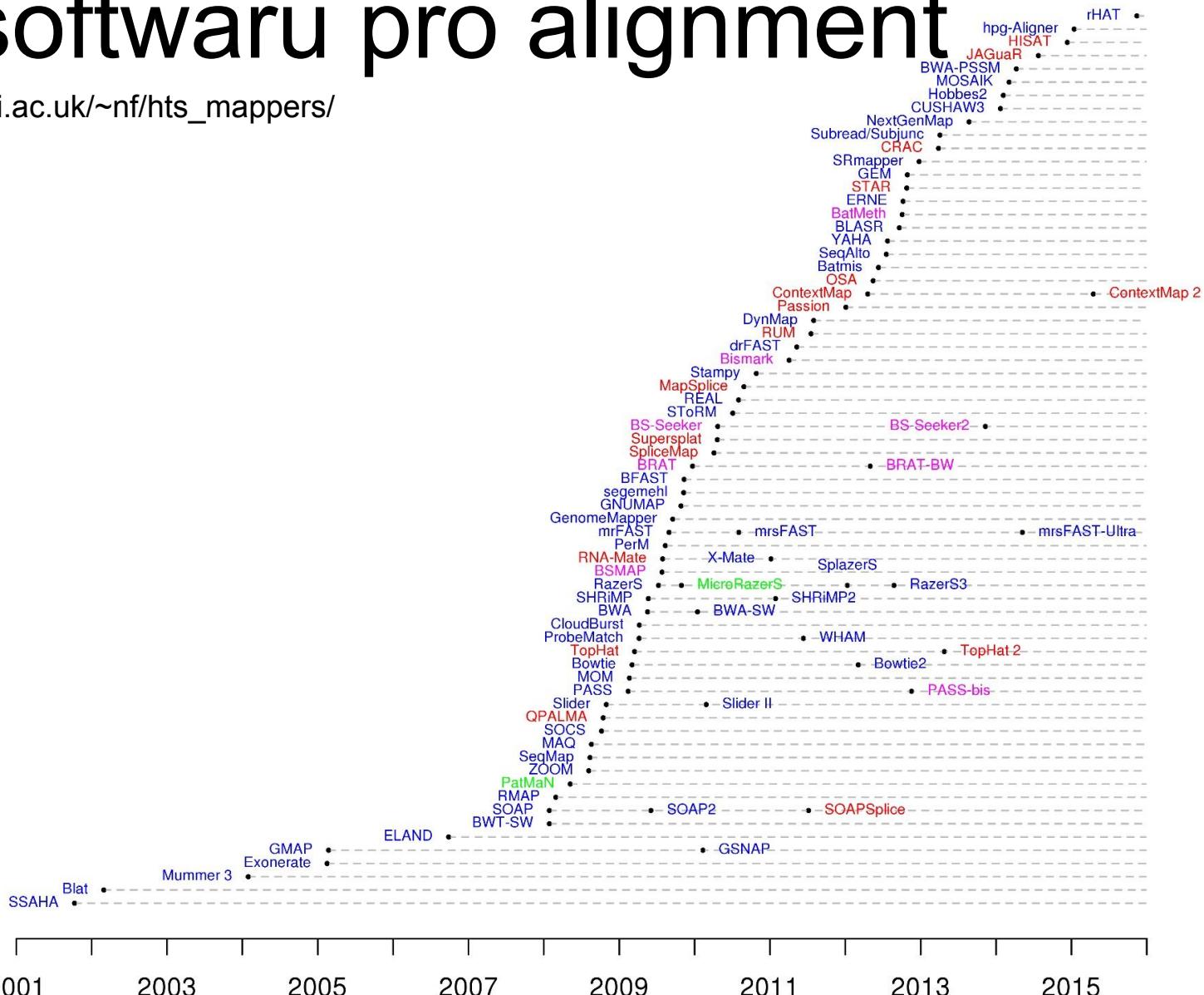
**Kolik clusterů jste obdrželi? Jsou některé KIR geny rozděleny do více clusterů?**

\*[https://usegalaxy.org/u/tomas\\_hron/h/kir-clustering-finished](https://usegalaxy.org/u/tomas_hron/h/kir-clustering-finished)



# Vývoj softwaru pro alignment

[http://www.ebi.ac.uk/~nf/hts\\_mappers/](http://www.ebi.ac.uk/~nf/hts_mappers/)



# BWA-MEM

- Je vylepšená verze Burrows-Wheeler Alignment algoritmu
- Optimalizován na čtení  $\geq 70\text{bp}$
- Je rychlejší než BWA a má lepší detekci variant, inserci a deleci
- Více informací na <http://bio-bwa.sourceforge.net/>



HPST, s.r.o.





Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NGS: Gemini

NGS: QC

NGS: Variant detection

NGS: Alignment

NGS: FASTA nad FASTQ processing

NGS: Variant annotation

NGS: SAMtools

NGS: BAMtools

NGS: BEDtools

NGS: VCF manipulation

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.12.1)

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below

Using reference genome

Human (hg19)

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

7: FASTQ Groomer on data 2

Specify dataset with forward reads

Select second set of reads

6: FASTQ Groomer on data 1

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-l; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

History

23: markDuplicates on data 20: MarkDuplicates BA M output

24: MarkDuplicates on data 20: MarkDuplicate met rics

21: Map with BWA-MEM on data 8 and data 9 (ma pped reads in BAM format)

20: Map with BWA-MEM on data 6 and data 7 (ma pped reads in BAM format)

Vyberte správnou verzi genomu

17: FastQC on data 9: Ra wData

16: FastQC on data 9: We bpage

15: FastQC on data 8: Ra wData

14: FastQC on data 8: We bpage

# Cvičení 3 - mapování DNA knihovny KIR receptorů

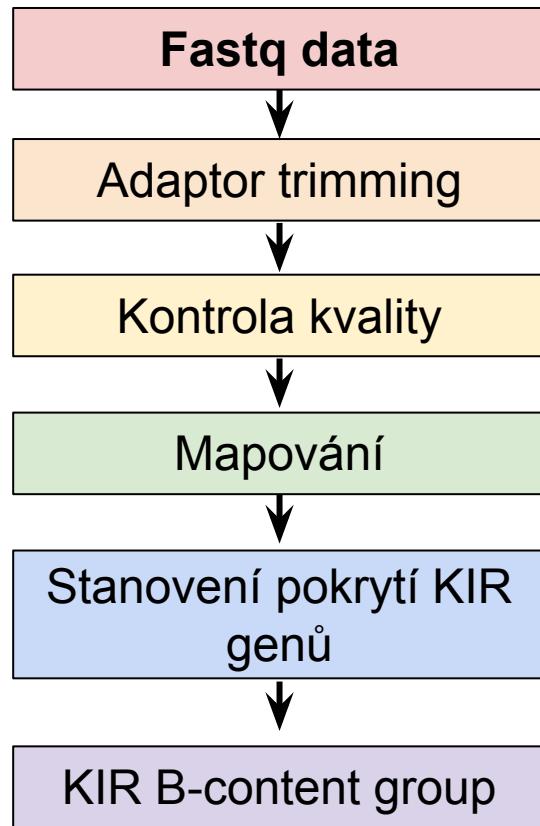
1. Importujte historii “**KIR-mapping**” ([https://usegalaxy.org/u/tomas\\_hron/h/kir-mapping](https://usegalaxy.org/u/tomas_hron/h/kir-mapping)).
2. Proveďte 3 různá mapování pomocí nástroje “**Map with BWA-MEM**” s referenční sekvencí:
  - i) *KIR\_gen-clustered*
  - ii) *KIR\_nuc-clustered*
  - iii) *Human: hg38 Canonical*
3. Pro každý BAM soubor spusťte nástroj “**Samtools flagstat**” a zjistěte kolik procent readů úspěšně zamapovalo

**Která reference se zdá být nevhodnější?**

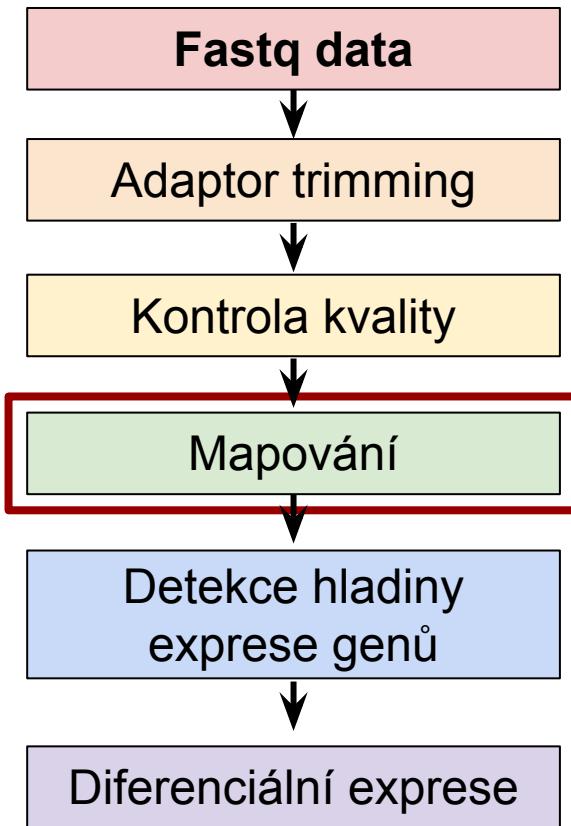
\*[https://usegalaxy.org/u/tomas\\_hron/h/kir-mapping-finished](https://usegalaxy.org/u/tomas_hron/h/kir-mapping-finished)



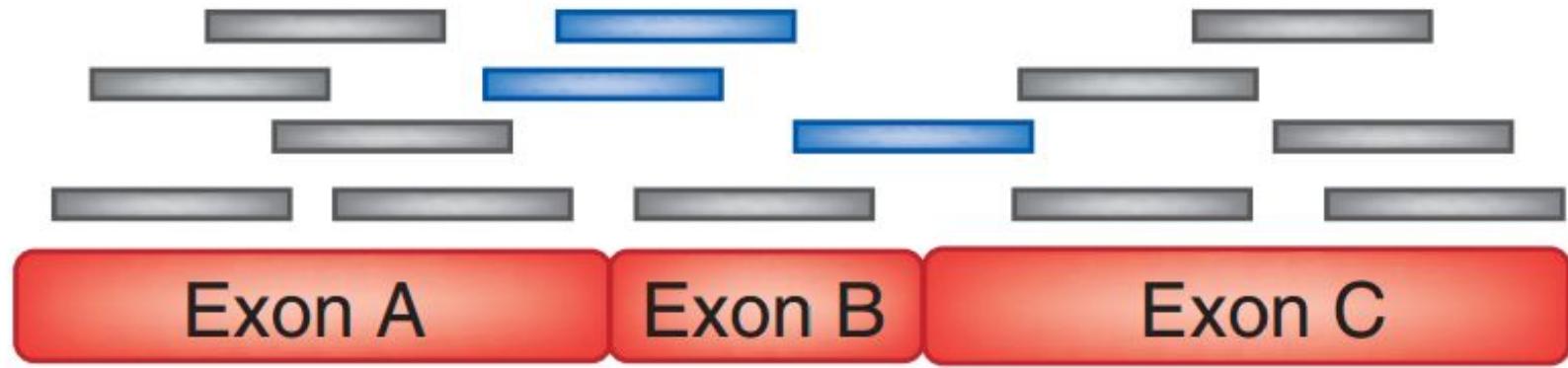
## Genotypizace KIRs



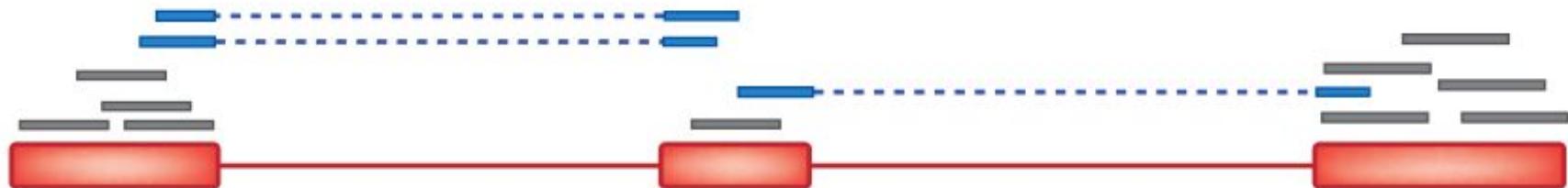
## Diferenciální exprese genů u D. melanogaster



# RNA mapping



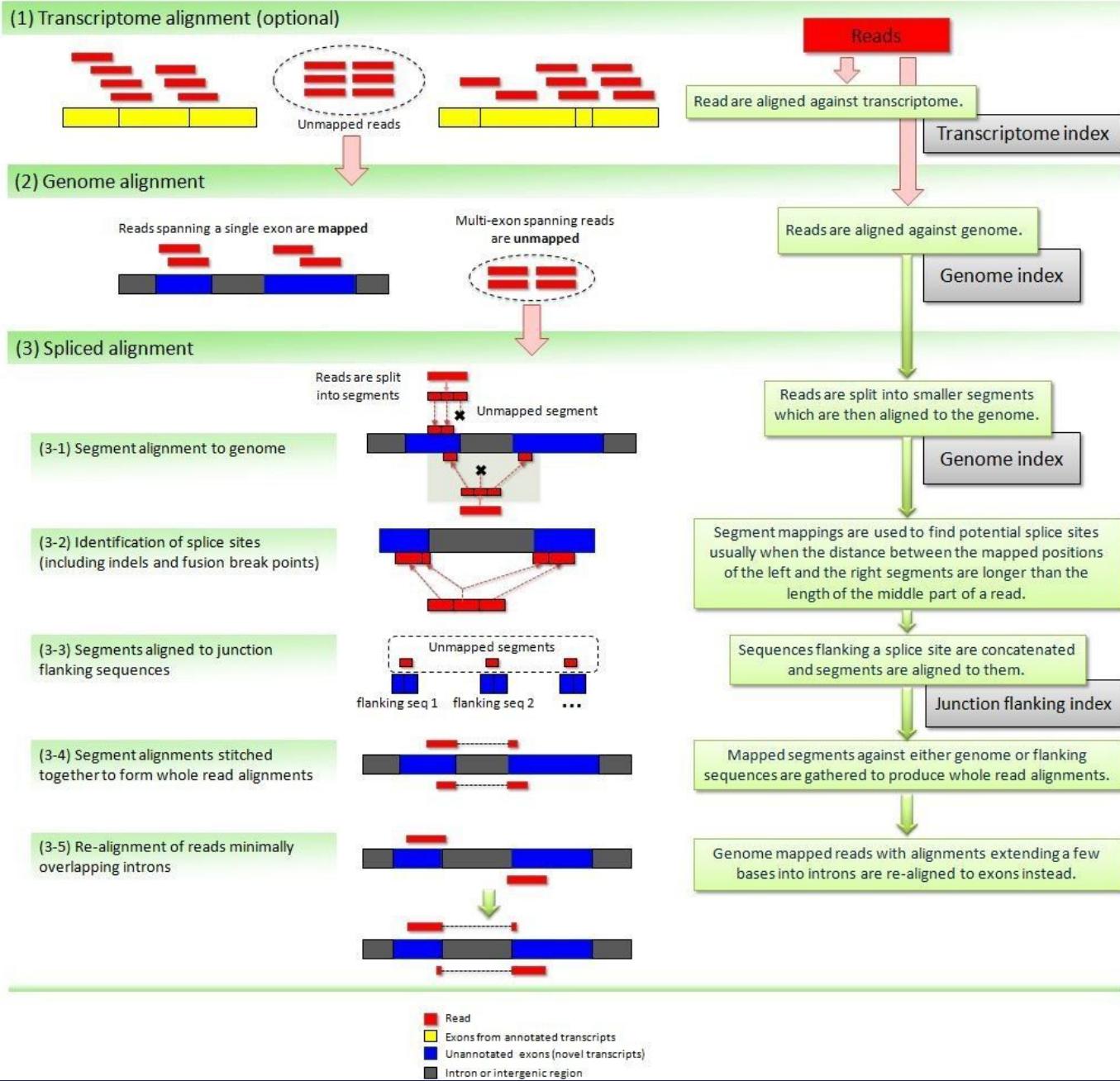
Processed mRNA



Mapping to genome

Trapnell a Salzberg, 2009

# TopHat2



# TopHat

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.0)

Versions Options

Is this single-end or paired-end data?

Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads

1: c1-f.fastq

Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads

2: c1-r.fastq

Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs

28

-r--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.

Std. Dev for Distance between Mate Pairs

20

--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

Report discordant pair alignments?

No

--no-discordant

Use a built-in reference genome or own from your history

Use a built-in genome

Built-ins genomes were created using default options

Select a reference genome

Fruit Fly (Drosophila melanogaster): dm3

If your genome of interest is not listed, contact the Galaxy team

TopHat settings to use

Use Defaults

You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?

No

Job Resource Parameters

Use default job resource parameters

Execute

History

search datasets

Unnamed history

4 shown

58.57 MB

4: c2-r.fastq

3: c2-f.fastq

2: c1-r.fastq

1: c1-f.fastq



HPST, s.r.o.



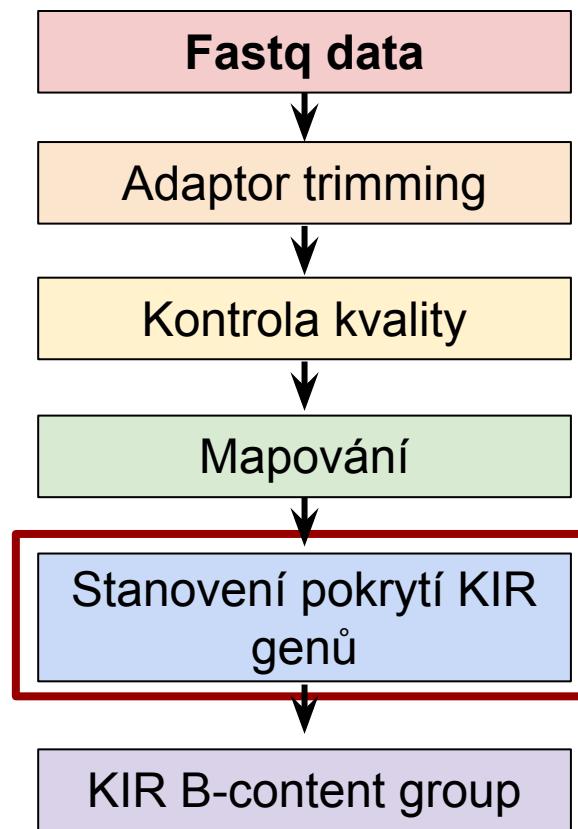
# Cvičení\_4 - mapování RNAseq

1. Importujte historii “***RNA-Mapping***” ([https://usegalaxy.org/u/tomas\\_hron/h/rna-mapping](https://usegalaxy.org/u/tomas_hron/h/rna-mapping)).
2. Proveďte mapování pomocí programu “***TopHat***”, nastavte parametry:
  - RNA-Seq FASTQ file, forward reads: c1-r1-f-x.fastq
  - RNA-Seq FASTQ file, reverse reads: c1-r1-r-x.fastq
  - Mean Inner Distance between Mate Pairs = 28
  - Use built-in genome
  - Select the reference genome = dm6
  - TopHat settings to use = Full parameter list (This is done to be able to specify the strandedness of the library)
  - Library Type = FR First Strand
3. **Odvodíte, co obsahují výsledné soubory?**
4. Otevřete BAM soubor s příponou “***accepted hits***” v IGV (<https://igv.org/app/>) a zobrazte oblast ***chrX:11,801,434-11,807,435***. **Dokážete najít čtení, která mapují přes dva introny (split reads)?**
5. Klikněte pravým tlačítkem do prostoru alignmentu v IGV a vyberte možnost ***Sashimi Plot***. **Domyslíte, co zobrazený graf ukazuje?**

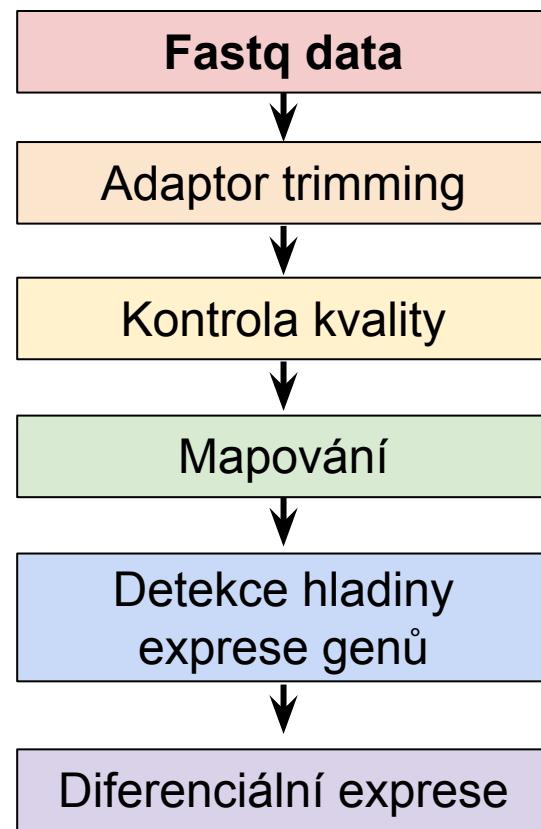
\*[https://usegalaxy.org/u/tomas\\_hron/h/rna-mapping-finished](https://usegalaxy.org/u/tomas_hron/h/rna-mapping-finished)



## Genotypizace KIRs



## Diferenciální exprese genů u *D. melanogaster*



# SAM FLAGS

#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment

<http://www.samformat.info/sam-format-flag>



# SAM FLAGS

Kombinace bitových FLAGů

Příklad: FLAG = 65 čtení je první v páru  
párová čtení ( $2^0 = 1$ ) + první v páru ( $2^8 = 64$ ) = 65

Vyzkoušejme si to zde:

<http://www.samformat.info/sam-format-flag>

# QualiMap BamQC

## statistika mapování

QualiMap BamQC (Galaxy Version 2.2.2d+galaxy1)

Favorite  Options

Mapped reads input dataset

No bam dataset available.

(-bam)

Reference genome regions to calculate mapping statistics for

All (whole genome)

Generate per-base coverage output

Yes  No

Produce additional tabular output listing the coverage at every site (omitting only zero-coverage positions) in the selected regions of the genome. Caution: Will generate a huge dataset for anything but small input genomes or restricted regions! (-oc)

Skip duplicate reads

Select/Unselect all

Reads flagged as duplicates in input  
 Duplicates detected by Qualimap

(--skip-dup-mode)

Settings affecting specific plots

Job Resource Parameters

Use default job resource parameters

Email notification

Yes  No



HPST, s.r.o.

 Agilent Technologies  
Autorizovaný distributor

# Cvičení 5 - Statistika mapování a Filtrování BAM souboru

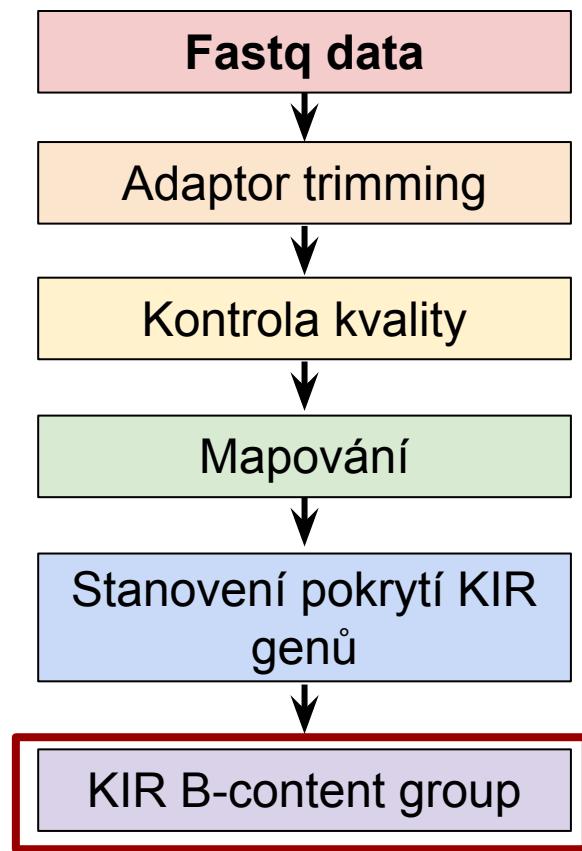
1. Importujte historii “**KIR-Genotyping**” ([https://usegalaxy.org/u/tomas\\_hron/h/kir-genotyping](https://usegalaxy.org/u/tomas_hron/h/kir-genotyping)).
2. Spusťte nástroj **Filter BAM datasets** a vyberte BAM soubor, v kterém jsou použity jako reference clusterované genové sekvence KIR. Odfiltrujte všechny ready, které nemapují k referenci unikátně (Mapping Quality = 0)
  - o pro filtrování použijte hodnotu MapQ  $\geq 1$
3. Pro původní i výsledný BAM soubor použijte nástroj **QualiMap BamQC**

**Kolik readů nám zbylo?**

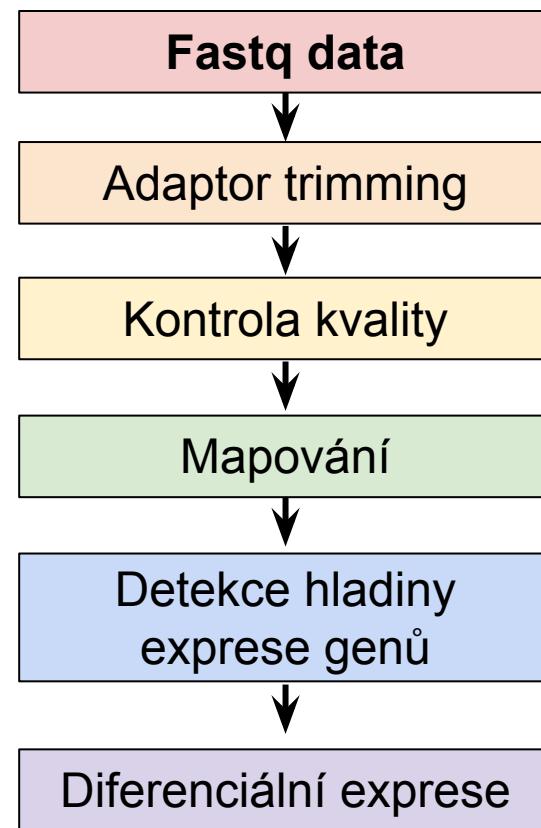
**Které KIR geny jsou sekvenačně pokryty unikátními ready?**

\*[https://usegalaxy.org/u/tomas\\_hron/h/kir-genotyping-finished](https://usegalaxy.org/u/tomas_hron/h/kir-genotyping-finished)

## Genotypizace KIRs



## Diferenciální exprese genů u D. melanogaster



# Cvičení 6 - Donor KIR B-content group calculator

Overview

IMGT/HLA

KIR

MHC

NHKIR

HPA

ESTDAB

Contact

Support

IPD / KIR / B-CONTENT

## Donor KIR B-content group calculator

Killer-cell Immunoglobulin-like Receptor (KIR) genes form a diverse, immunogenetic system unlinked to HLA. Group A and B KIR haplotypes have distinctive centromeric (Cen) and telomeric (Tel) gene-content motifs. With the goal of developing a donor selection strategy to improve transplant outcome, Cooley *et al.*, compared the contribution of these motifs to the clinical benefit conferred by B haplotype donors. Donor KIR genotype influenced transplantation outcome for AML, but not ALL, after HLA-matched or mismatched T-cell replete unrelated donor transplants. Compared to A haplotype motifs, centromeric and telomeric B motifs both contributed to relapse protection and improved survival, but Cen-B homozygosity had the strongest independent effect. KIR genotyping several best HLA-matched potential donors should substantially increase the frequency of transplants using unrelated donor grafts with favorable KIR gene content. Adopting this practice could result in superior disease-free survival for patients transplanted for AML.

Disclaimer - This research tool is being offered as a tool to predict donor KIR B-content groups assignments as reported in:

S Cooley, DJ Weisdorf, LA Guethlein, JP Klein, T Wang, CT Le, SGE Marsh, D Geraghty, S Spellman, MD Haagenson, M Ladner, E Trachtenberg, P Parham and JS Miller.

Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia.

Blood (2010) 116:2411-9.

No information entered into this tool is collected or stored on our servers.

This calculator allows you to enter the genotypes for up to five prospective donors, and receive their assignments to one of 3 groups based on KIR B-content. The groups, "Neutral", "Better", "Best", refer to the associated relapse protection seen in T-cell replete URD HCT for AML. Simply enter the prospective donor's KIR genotype by selecting the appropriate box for each gene. A tick indicates presence of a gene in the donor's genotype.

## Resources

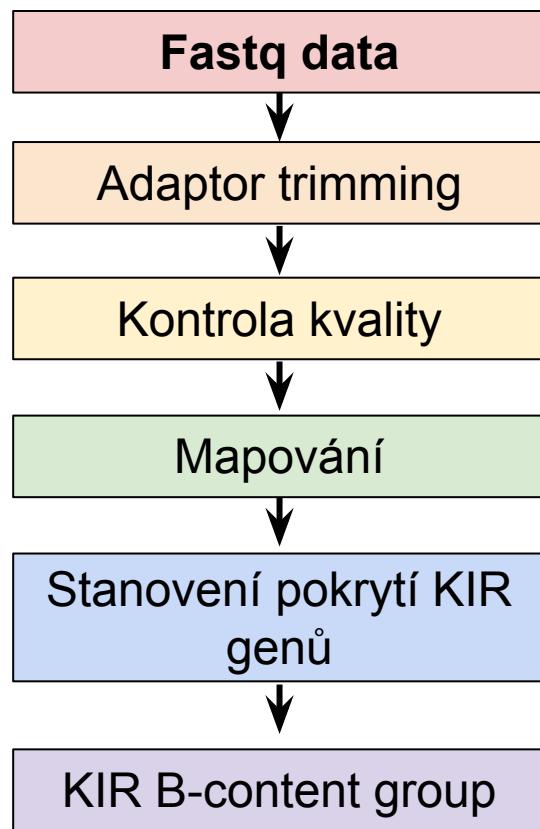
- [!\[\]\(1dfe18043d832d229b8b701f5ceb86f3\_img.jpg\) About >](#)
- [!\[\]\(deaa5a45331e146c9936b15bdf00380b\_img.jpg\) Statistics >](#)
- [!\[\]\(ace37c88c2fab8d06ef748934c907dc6\_img.jpg\) Publications >](#)
- [!\[\]\(e2d1094ec38a73c7083b3270c6ad1448\_img.jpg\) Nomenclature >](#)
- [!\[\]\(6c2411fa50d1a1817eab904bca235e26\_img.jpg\) Genes >](#)
- [!\[\]\(ab6d8b92aef2d3ba76acedecafa0d90c\_img.jpg\) Alleles >](#)
- [!\[\]\(8c86da806ef5ddcfd6e79d8d31e051c3\_img.jpg\) Haplotypes >](#)
- [!\[\]\(efe6cfcea7289a89b29dda3461c12da7\_img.jpg\) Genotypes >](#)
- [!\[\]\(9470e27b9bc4a6c6af9fb878b16cb776\_img.jpg\) Alignments >](#)
- [!\[\]\(19f600d03d58a444f936c99b6a277d03\_img.jpg\) Releases >](#)
- [!\[\]\(d77da847fd22a437d366a1497b14898e\_img.jpg\) BLAST >](#)
- [!\[\]\(f5eb1786eef62a8c70247d84755a04b7\_img.jpg\) B-Content >](#)
- [!\[\]\(b586ccf7e6a6579a4a111770d317b79f\_img.jpg\) Ligands >](#)
- [!\[\]\(e84f3322035624a0326e44e4f074ea9d\_img.jpg\) FTP >](#)
- [!\[\]\(8faf7802d2101de5341d0d385d66d1bd\_img.jpg\) GitRepos >](#)
- [!\[\]\(d203e2c628a740cc38aed2e903bebd9b\_img.jpg\) Cells >](#)
- [!\[\]\(f42ab9456c1b44c2ce89d235927568ee\_img.jpg\) FAQ >](#)
- [!\[\]\(d11741f7f62c8b21afb4a656ffe53c74\_img.jpg\) Submissions >](#)



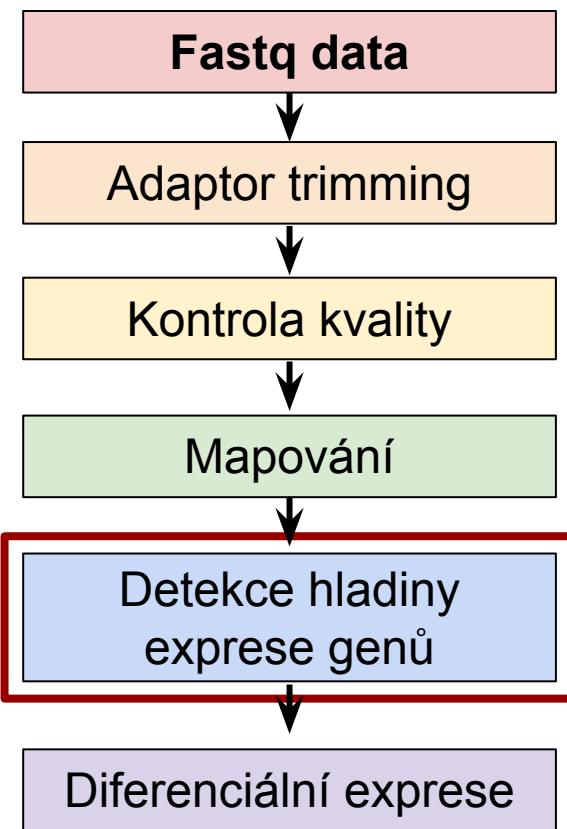
HPST, s.r.o.



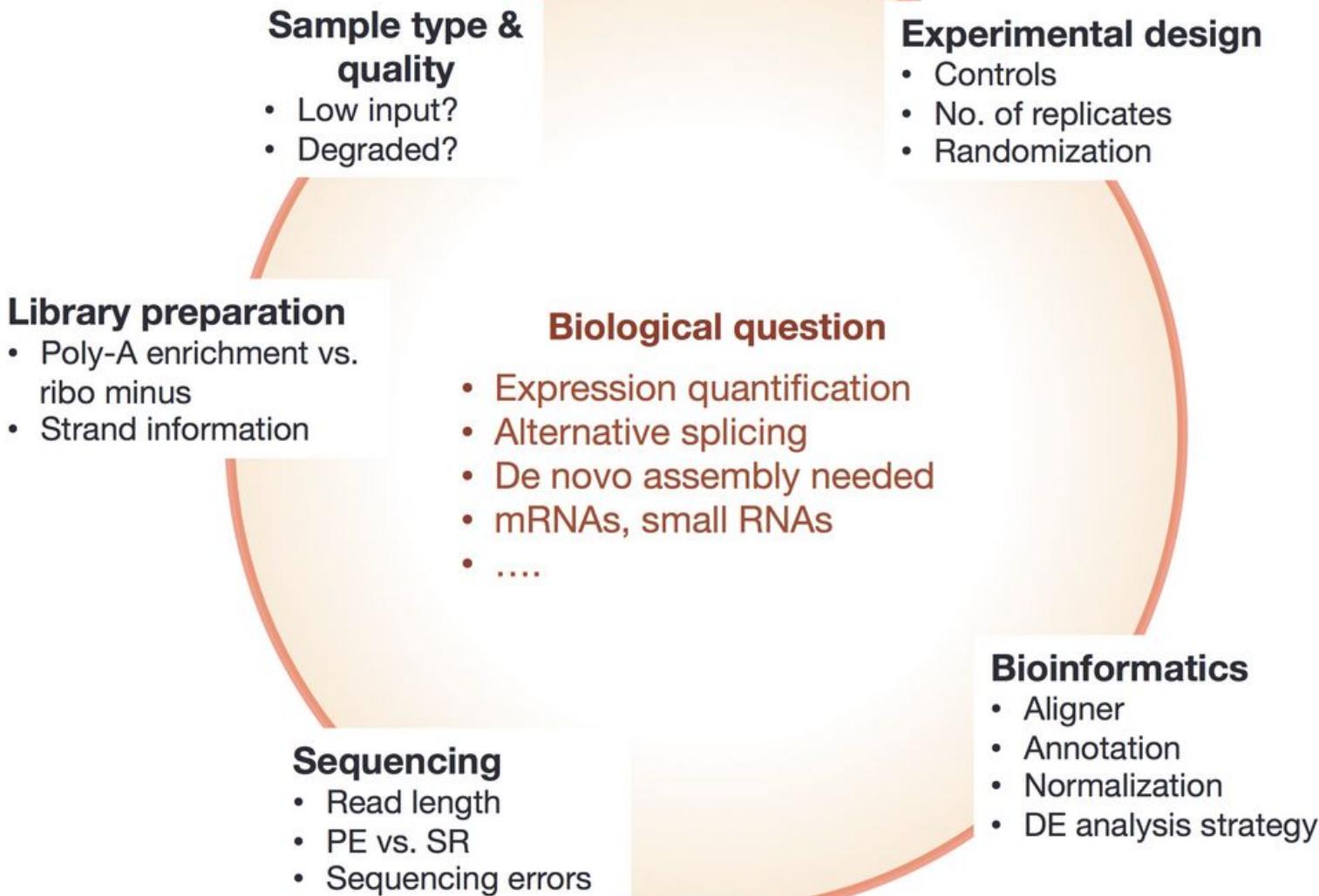
## Genotypizace KIRs

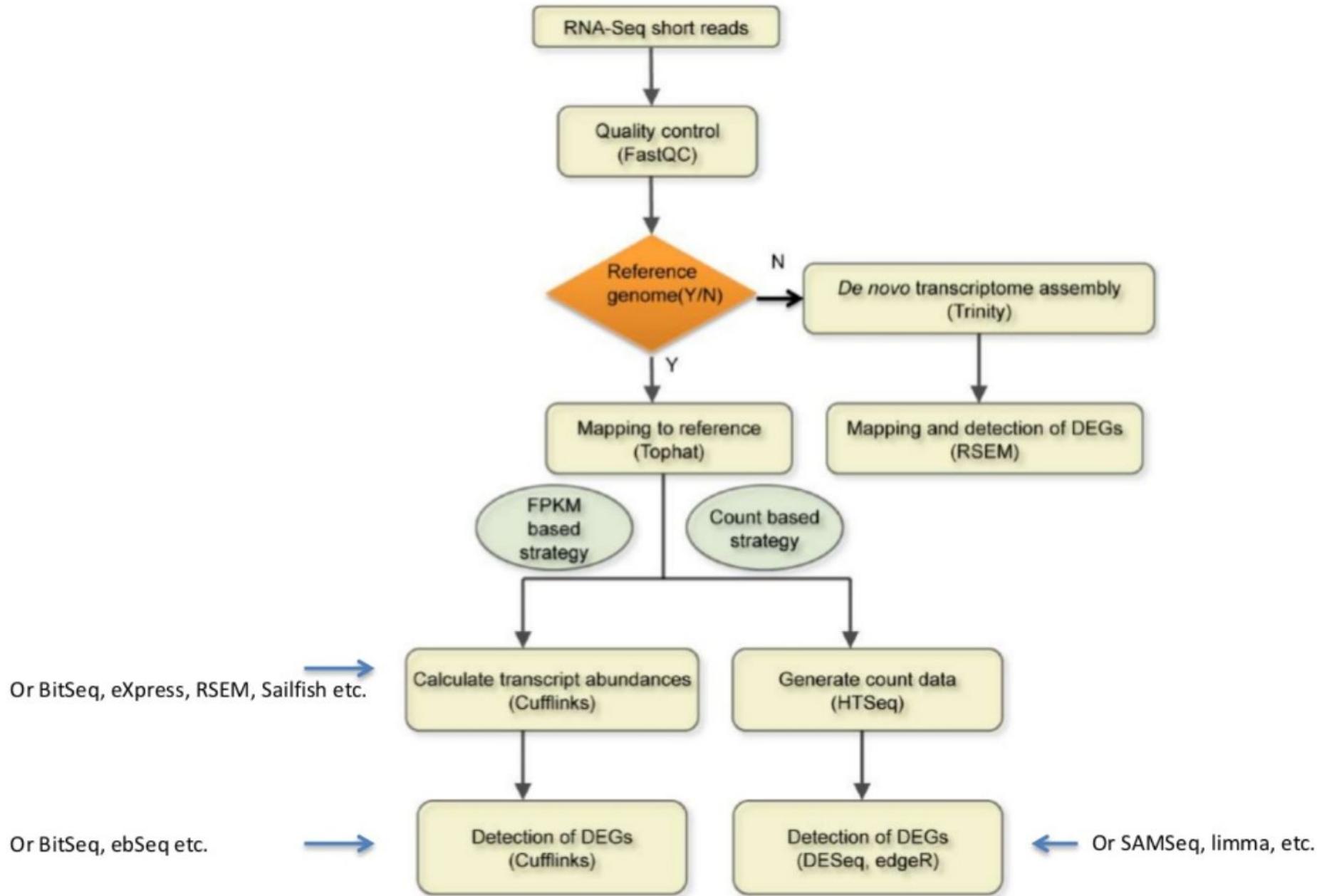


## Diferenciální exprese genů u D. melanogaster



# Everything's connected...

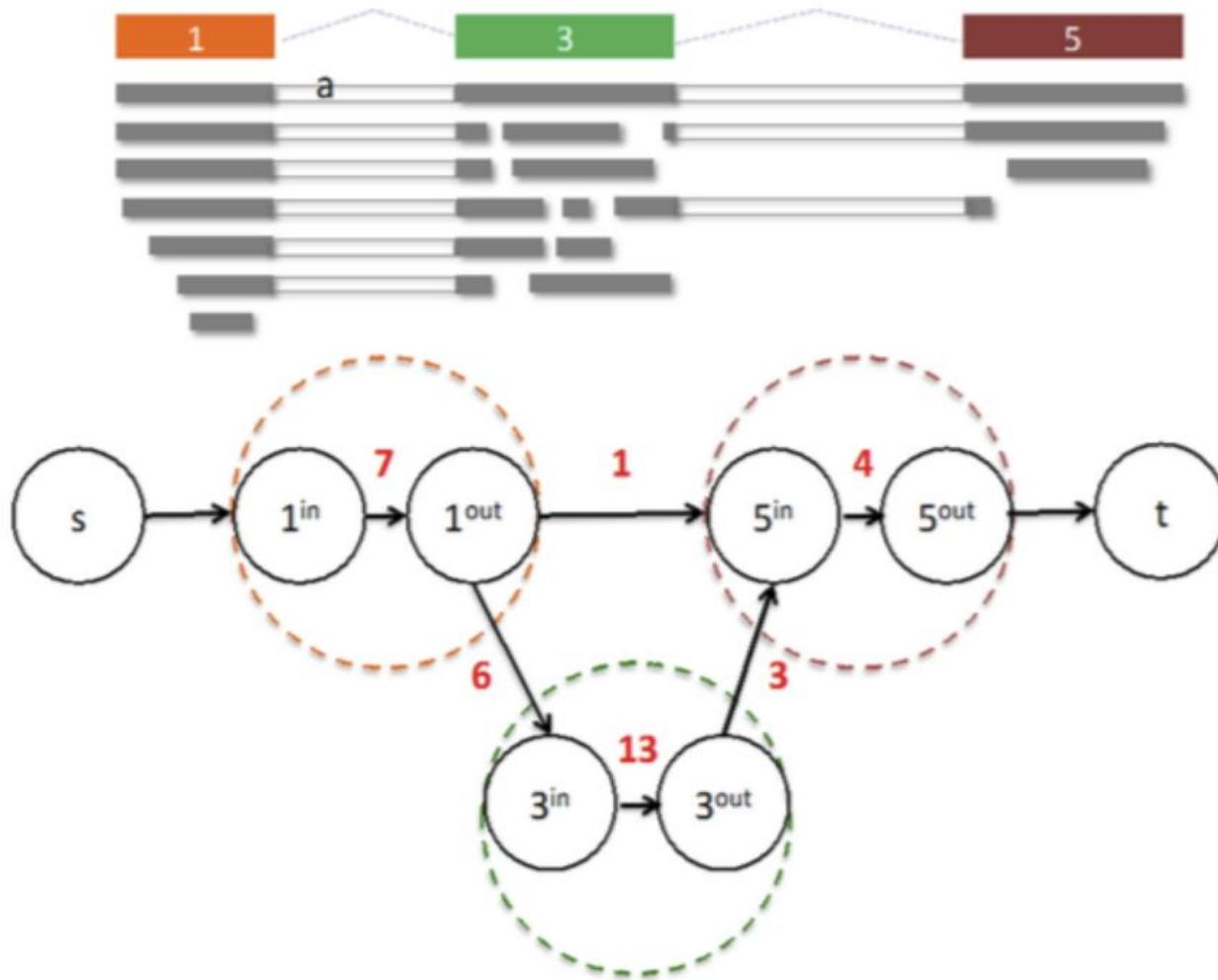




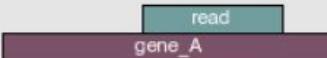
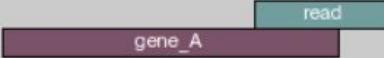
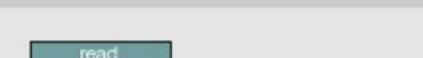
# From reads to RNA molecules

- **Reference genome-based** - an assembled genome exists for a species for which an RNAseq experiment is performed. It allows reads to be aligned against the reference genome and significantly improves our ability to reconstruct transcripts. This category would obviously include humans and most model organisms.
- **Reference genome-free** - no genome assembly for the species of interest is available. In this case one would need to assemble the reads into transcripts using *de novo* approaches.

# Transcript counting (StringTie)



# Gene counting (HTseq-count)

	union	intersection _strict	intersection _nonempty
 A single read (green) overlaps a single gene (purple). The read starts within the gene and ends outside.	gene_A	gene_A	gene_A
 A single read (green) starts within a gene (purple) and ends outside. It does not cover the entire gene.	gene_A	no_feature	gene_A
 A single read (green) spans two adjacent genes (purple and brown). It starts within the first gene and ends within the second.	gene_A	no_feature	gene_A
 Two reads (green) overlap two adjacent genes (purple and brown). Both reads start within the first gene and end within the second.	gene_A	gene_A	gene_A
 A single read (green) overlaps two adjacent genes (purple and blue). The read starts within the first gene and ends within the second.	gene_A	gene_A	gene_A
 A single read (green) is positioned such that it could map to either of two adjacent genes (purple and blue), starting within one and ending within the other.	ambiguous	gene_A	gene_A
 A single read (green) is positioned such that it could map to either of two adjacent genes (purple and blue), starting within one and ending within the other.	ambiguous	ambiguous	ambiguous

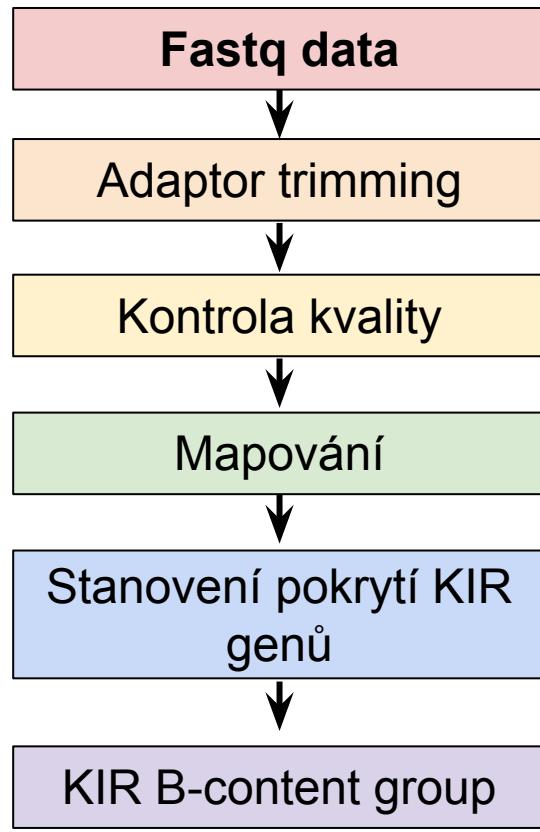
# Cvičení 7 - Stanovení genové exprese

1. Importujte historii "RNA-DEA" ([https://usegalaxy.org/u/tomas\\_hron/h/rna-dea](https://usegalaxy.org/u/tomas_hron/h/rna-dea)), kde naleznete 6 BAM souborů (dvě experimentální podmínky v triplikátu)  
*Condition1\_replicate1, Condition1\_replicate2, Condition1\_replicate3*  
*Condition2\_replicate1, Condition2\_replicate2, Condition2\_replicate3*
2. Vytvořte tabulku počtu čtení pro jednotlivé geny pomocí nástroje "**htseq-count**". Vyberte všechny BAM soubory, zvolte "model=UNION", "Strand=yes", "Minimal align quality=10".

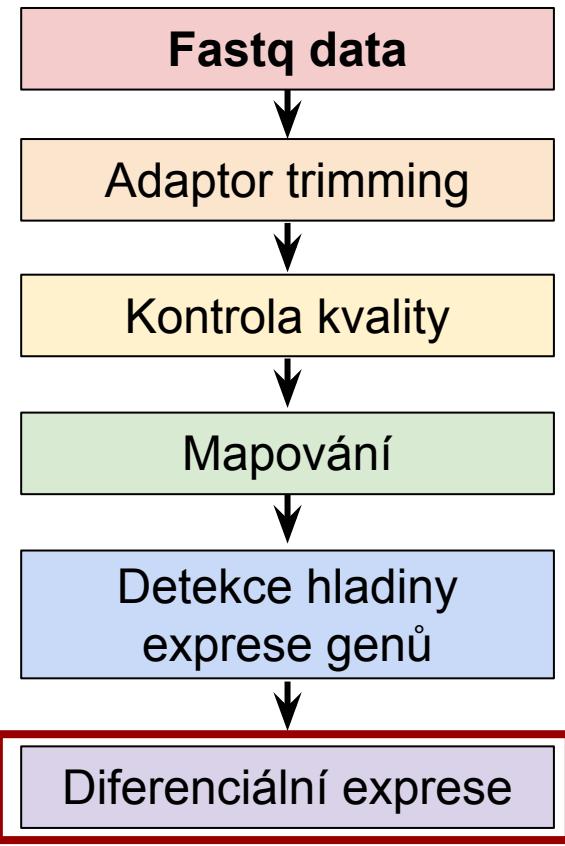
\*[https://usegalaxy.org/u/tomas\\_hron/h/rna-dea-finished](https://usegalaxy.org/u/tomas_hron/h/rna-dea-finished)



## Genotypizace KIRs



## Diferenciální exprese genů u D. melanogaster



# Read count normalisation (within sample)

Table 11: Normalization methods for the comparison of gene read counts within the same sample.

Name	Details	Comment
RPKM (reads per kilobase of exons per million mapped reads)	<ol style="list-style-type: none"><li>For each gene, count the number of reads mapping to it.</li><li>Divide that count by: the length of the gene in base pairs divided by 1,000 multiplied by the total number of mapped reads divided by <math>10^6</math>.</li></ol> $RPKM_i = \frac{\text{read count of gene } i}{\left(\frac{\text{length of gene } i}{10^3}\right)\left(\frac{\text{library size}}{10^6}\right)}$	<ul style="list-style-type: none"><li>introduces a bias in the per-gene variances, in particular for lowly expressed genes (Oshlack and Wakefield, 2009)</li><li>implemented in edgeR's <code>rpkms()</code> function</li></ul>
FPKM (fragments per kilobase...)	<ol style="list-style-type: none"><li>Same as RPKM, but for paired-end reads:</li><li>The number of fragments (defined by two reads each) is used.</li></ol>	<ul style="list-style-type: none"><li>implemented in DESeq2's <code>fpkm()</code> function</li></ul>
TPM	<p>Instead of normalizing to the total library size, TPM represents the abundance of an individual gene <math>i</math> in relation to the abundances of the other transcripts (e.g., <math>j</math>) in the sample.</p> <ol style="list-style-type: none"><li>For each gene, count the number of reads mapping to it and divide by its length in base pairs (= counts per base).</li><li>Multiply that value by 1 divided by the sum of all counts per base of every gene.</li><li>Multiply that number by <math>10^6</math>.</li></ol>	<ul style="list-style-type: none"><li>details in Wagner et al. (2012)</li></ul>

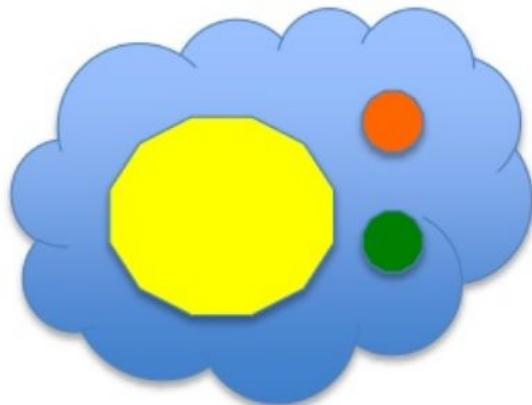
$$TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}}$$

# TMM – Trimmed Mean of M values

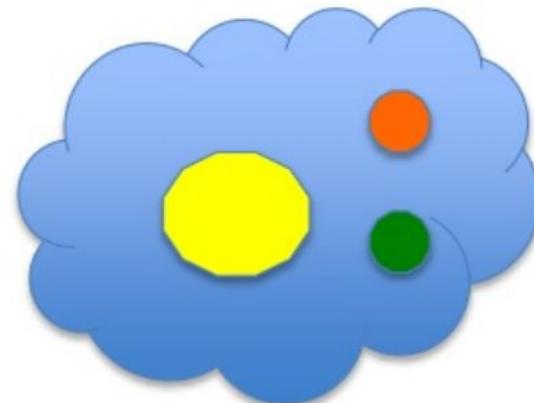
Attempts to correct for differences in RNA *composition* between samples

E.g. if certain genes are very highly expressed in one tissue but not another, there will be less “sequencing real estate” left for the less expressed genes in that tissue and RPKM normalization (or similar) will give biased expression values for them compared to the other sample

RNA population 1



RNA population 2



Equal sequencing depth -> orange and red will get lower RPKM in RNA population 1 although the expression levels are actually the same in populations 1 and 2

Robinson and Oshlack Genome Biology 2010, 11:R25, <http://genomebiology.com/2010/11/3/R25>

# Differential expression analysis - DEseq

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.1.8.3) ▼ Options

**Factor**

**1: Factor**

**Specify a factor name**

Conditions

Only letters, numbers and underscores will be retained in this field

**Factor level**

**1: Factor level**

**Specify a factor level**

Condition 1

Only letters, numbers and underscores will be retained in this field

**Counts file(s)**

84: htseq-count on collection 37

**2: Factor level**

**Specify a factor level**

Condition 2

Only letters, numbers and underscores will be retained in this field

**Counts file(s)**

92: htseq-count on collection 57

**+ Insert Factor level**

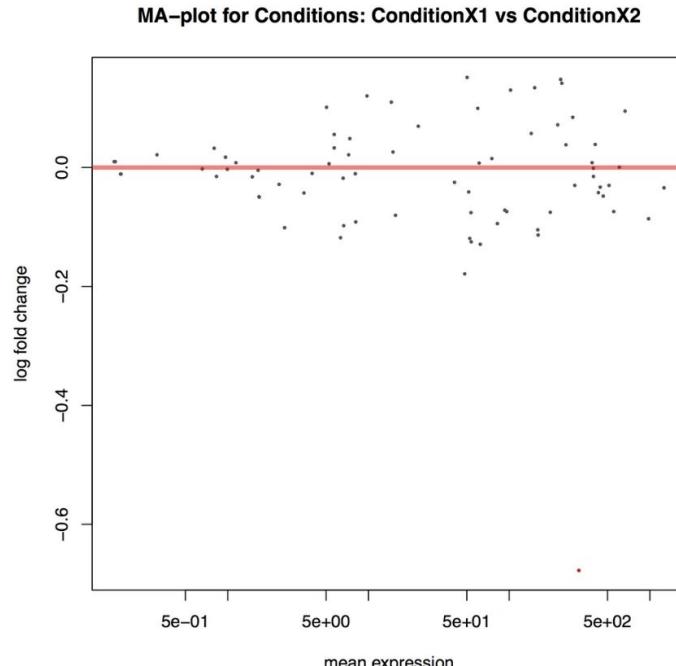
**+ Insert Factor**

**Visualising the analysis results**

Yes No

This will produce [output](#) as shown below. The columns are: (1) gene identifier, (2) mean normalised counts, averaged over all samples from both conditions, (3) logarithm (base 2) of the fold change, (4) the standard error estimate for the log2 fold change estimate, (5) [Wald test](#) statistic, (6) p value for the statistical significance of this change, and (7) *p*-value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate ([FDR](#)). There is only one gene with significant change in gene expression between conditions: **CG1803-RC** with *p*-value =  $1.6 \times 10^{-5}$

1	2	3	4	5	6	7
CG1803-RC	313.399854993883	-0.677347951975255	0.130547283480314	-5.18852582694606	2.11965367545954e-07	1.61093679334925e-05
CG10352-RA	3.97546094727023	-0.00977069078261261	0.0747042077056103	-0.130791706152836	0.895940084165715	0.981157785658315
CG10353-RD	22.6113188342966	0.0696827956291263	0.129706150655657	0.537235861806736	0.591104702591145	0.981157785658315
CG10362-RA	50.2265915083909	0.151864019842212	0.137485127753392	1.10458507275501	0.26933942115096	0.981157785658315



# Cvičení 8 - Analýza diferenciální exprese

1. Otevřete historii "RNA-DEA" ([https://usegalaxy.org/u/tomas\\_hron/h/rna-dea](https://usegalaxy.org/u/tomas_hron/h/rna-dea)).
2. Proveďte analýzu diferenciální exprese pomocí "**DESeq2**" (změňte na verzi 2.1.8.3). Data obsahují dva vzorky ve třech replikátech. Nastavte název faktoru = "Condition" a vytvořte dvě úrovně faktoru (Factor level) - Condition1 a Condition2. Do těchto úrovní vyplňte příslušná data. "Visualising analysis results=yes", "Output all levels vs all levels=no".
3. Ve výsledné tabulce jsou hodnoty pro jednotlivé proměnné: (1) gene identifier, (2) mean normalised counts, averaged over all samples, (3) log2 of the fold change, (4) the standard error estimate for the log2 fold change, (5) [Wald test](#) statistic, (6) p value for the statistical significance of this change, (7) *p*-value adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate ([FDR](#)).
4. **Je nějaký gen ve vzorcích diferenciálně exprimován (signifikantně)?**

# RNAseq analysis tutorial

## RNAseq: Reference-based

This tutorial is inspired by an exceptional [RNAseq course](#) at the Weill Cornell Medical College compiled by Friederike Dündar, Luce Skrabanek, and Paul Zumbo and by tutorials produced by Björn Grüning (@bgruening) for Freiburg Galaxy instance. Much of Galaxy-related features described in this section have been developed by Björn Grüning (@bgruening) and configured by Dave Bouvier (@davebx).

- <https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/rb-rnaseq/tutorial.html>
- DATA:  
usegalaxy.org:  
Shared\_Data/Data\_Libraries/Tutorials/RNA\_seq\_(RB)





## Autorizovaný distributor Agilent Technologies



Produkty  
Služby

Kontakty  
Servisní požadavek

O nás  
Kariéra

Knihovna Agilent  
Odkazy

Novinky

Semináře

21. 3. 2018 Novinky a trendy Agilent Technologies 2018

Workshop

23.1.-24.1. 2018 Analýza dat z NGS s Galaxy bez znalosti příkazového rádku

Nové produkty

3D tomografický fluorescenční mikroskop od Nanolive

Nové aplikace

ICP-QQQ - komplexní online knihovna aplikací

Školení

MassHunter pro GC/MS

OpenLAB CDS 2.x

LC školení, uživatelské dovednosti a diagnostika

GC školení, uživatelské dovednosti a diagnostika

NOVINKA: GC školení, uživatelské dovednosti a diagnostika pro pokročilé uživatele

MS (single-quad/QQQ pro GC) školení,  
uživatelské dovednosti a diagnostika

hpst.cz

**Novinky a trendy  
Agilent  
Technologies**  
**21. 3. 2018**  
**Přihlaste se!**

Nejžhavější novinky ze světa Agilent na jednom místě!

**FOOD TESTING INSIGHTS**  
  
Budte v obraze v oboru kontroly kvality a autenticity potravin!

  
**TRUSTED  
ANSWER.  
TOGETHER.**  
  
DAKO - Komplexní řešení pro patologické laboratoře

**MOLECULAR SPECTROSCOPY  
WEBINAR SERIES**  


**STRATEGIES FOR GC SUCCESS**  
**AGILENT INTUVO 9000 GC**  




[thron@hpst.cz](mailto:thron@hpst.cz)  
[dqq@hpst.cz](mailto:dqq@hpst.cz)



HPST, s.r.o.

 Agilent Technologies  
Autorizovaný distributor